21782

NATIONAL LIBRARY
OTTAWA

CANADA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR... Patsy Suk Mui Chan .....

TITLE OF THESIS... The Multicollinearity Problem
... in Regression Analysis .....

.....

UNIVERSITY..... University of Alberta .....

DEGREE FOR WHICH THESIS WAS PRESENTED.... M. Sc. .....

YEAR THIS DEGREE GRANTED... 1974 .....

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed). Patsy Suk Mui Chan

PERMANENT ADDRESS:

11028 - 84 Ave.,
Edmonton, Alberta

.....

DATED.. 24 May 19

NL-91 (10-68)

THE UNIVERSITY OF ALBERTA

THE MULTICOLLINEARITY PROBLEM IN REGRESSION ANALYSIS

by

© PATSY S.M. CHAN

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

STATISTICS

DEPARTMENT OF MATHEMATICS

EDMONTON, ALBERTA

FALL, 1974

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled .....

...THE... MULTICOLLINEARITY.... PROBLEM....IN.....

....REGRESSION.....ANALYSIS.....

submitted by .....PATSY.....S.M.....CHAN.....

in partial fulfilment of the requirements for the degree of

Master of .....STATS.....

...Andy...Fenn.....
Supervisor

...K...L...McRae...

.....I...Basu.....

Date MAY...24...1974

ABSTRACT

The purpose of this study is to review the problem of multicollinearity in regression analysis. Specifically, the difficulties that arise when multicollinearity is present, the alternative procedures available for detecting the problem, and the methods by which it may be resolved, are discussed. It is discovered that an error exists in Kirchdorfer's method of detection, thus rendering the procedure invalid.

In considering the strengths and weaknesses of each method of detection or resolution, what crystallizes is the view that to date, no highly satisfactory means of treating the problem has yet appeared. An illustrative application of the theory is obtained using Farrar-Glauber's techniques to detect multicollinearity in a sample of economic data. Hoerl and Kennard's Ridge Trace is then constructed, followed by calculation of Mayer-Willke shrunken estimator to remedy the detected multicollinearity.

ACKNOWLEDGMENTS

I am deeply grateful to Dr. Feuerverger for his patient guidance and valuable suggestions throughout the preparation of this thesis, and to Mrs. Billie Chiang for her excellent typing.

TABLE OF CONTENTS

Chapter		Page
I	INTRODUCTION	1
	1.1 The Regression Model	1
	1.2 The Multicollinearity Problem	6
II	DETECTION OF MULTICOLLINEARITY	11
	2.1 Measures of Multicollinearity	11
	2.2 Frisch Bunch Map Analysis	15
	2.3 Tintner Method	16
	2.4 Farrar-Glauber Technique	21
	2.5 Kirchdorfer Procedure	24
III	RESOLVING MULTICOLLINEARITY	28
	3.1 Generalized Inverses	28
	3.2 Using Additional Data	33
	3.3 Incorporating Extraneous Information	35
	3.4 The Mean Square Error Criterion	42
	3.5 Principal Component Estimators	47
	3.6 Factor Analysis	55
	3.7 Ridge Regression	59
	3.8 Marquardt Generalized Inverse Estimators	62
	3.9 Mayer-Willke Shrunken Estimators	64
	3.10 Multicollinearity in Two-stage Least-squares	65
IV	AN EMPIRICAL STUDY OF MULTICOLLINEARITY	69
	4.1 Description of the Model	69
	4.2 Application of Farrar-Glauber Technique	71
	4.3 Ridge Analysis of the Data	73
	4.4 Calculation of Mayer-Willke Shrunken Estimator	76
	REFERENCES	78
	APPENDIX I	84
	APPENDIX II	89

LIST OF TABLES

Table	Description	Page
I	Imports, Production, Stock Formation and Consumption in France (in Milliards of New Francs at 1956 Prices)	8
II	Ratio of Original to Revised Estimates	10
III	Adjoints for Subsets and Full Set of Variables in the French Economy Data	18
IV	Eigenvalues of Correlation Matrix for the French Economy Data	53
V	Normalized Eigenvectors for the First Two Components	53
VI	Partial Correlation Coefficient c_{ij}. and associated t_{ij}. between Pairs of Variables with $R^2_{X_i}$ on Diagonal	72

LIST OF FIGURES

Figure		Page
1.	Bunch Maps for Variables in the French Economy Data	17
2.	Ridge Trace for Inflation Data	74
3.	Squared Length of Coefficient Vector	75

CHAPTER I

INTRODUCTION

One of the most vexing problems in multiple regression analysis is that of multicollinearity, a term used to denote the presence of near linear relationships among the "independent" variables. Although econometricians and others seldom face the situation in which there is perfect multicollinearity, that is, one or more variables are exact linear combinations of other variable(s), high intercorrelation is nevertheless often an inevitable occurrence. This is due to the fact that economic variables are not generated by experimentally controlled conditions. It would therefore be of considerable value to investigate the problem of multicollinearity and the difficulties associated with a multicollinear set of data. Such was the intent of this study.

1.1 The Regression Model

The model on which our discussion centres is the familiar linear multiple regression equation

$$y = X\beta + u \qquad (1)$$

where X is an $n \times m$ matrix of n observations on m "independent" variables, rank $X = r \leq m < n$, β is an $m \times 1$ vector of unknown parameters and u is a vector of disturbances.

The minimal assumptions underlying the least squares theory are as follows:

> the elements of u are independently distributed random variables with mean zero and constant variance σ^2 .

According to the theory of least squares, we minimize $(y-X\beta)'(y-X\beta)$ and obtain the normal equations:

$$X'X\hat{\beta} = X'y \ . \tag{2}$$

Two cases can be considered depending on the singularity or nonsingularity of $X'X$.

(I) If X is of full rank, $(X'X)^{-1}$ exists and the least squares estimator is given by

$$\hat{\beta} = (X'X)^{-1}X'y \ .$$

The estimates $\hat{\beta}_i$ are unbiased, efficient and consistent as stated in the Gauss Markov theorem, a simplified proof of which is presented below.

The Gauss Markov Theorem. In the classical linear regression model, the best linear unbiased estimator of β is the least squares vector

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Proof (Plackett [44]). Let Wy be any unbiased estimator of β , i.e. $EWy = \beta$. Since

$$Ey = X\beta \ ,$$

this implies

$$WX = I \ .$$

Thus we can write

$$(X'X)^{-1} = WX(X'X)^{-1}$$

and obtain the identity

$$WW' = [(X'X)^{-1}X'][(X'X)^{-1}X']' + [W-(X'X)^{-1}X'][W-(X'X)^{-1}X']' \ .$$

That is, the diagonal elements of WW' are least when $W = (X'X)^{-1}X'$ which is the solution provided by least squares.

Adding to (1) the assumption that u is normally distributed, the results for the classical least squares model carry over. $\hat{\beta}$ has the same mean and variance as before. $(n-m)\hat{u}'\hat{u}/\sigma^2$ is distributed as χ^2_{n-m}. In addition, $\hat{\beta} = (X'X)^{-1}X'y$ is now normally distributed since it is a linear form in a normally distributed vector. It is also a uniformly minimum variance unbiased estimator of β.

Since the likelihood function of the sample is

$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{(2\sigma^2)}(y-X\beta)'(y-X\beta)} \ ,$$

maximizing it with respect to β is equivalent to choosing β such that $(y-X\beta)'(y-X\beta)$ is minimized. As this is precisely the least squares criterion established earlier, the maximum likelihood estimator is simply the least squares estimator. β is thus consistent and asymptotically efficient. The maximum likelihood estimator of σ^2 can be obtained as $(\hat{u}'\hat{u})/n$.

(II) If rank $X = r < m$, no unbiased estimator of β exists. However, a class of linear functions $L'\beta$, where L is an $m \times 1$ vector, may have unbiased estimators. These are the so-called "estimable functions". The estimable functionals L' are characterized by the property that a solution to the equations $a'X = L'$ exists, i.e. they are vectors in the row space of X. We have the following result.

Theorem 1.1.1. The best linear unbiased estimator of $L'\beta$ is $L'\hat{\beta}$, where $\hat{\beta}$ is any solution to the normal equations (2).

$L'\hat{\beta}$ may be expressed as $L'(X'X)^{g}X'y$ where $(X'X)^{g}$ is any generalized inverse of $(X'X)$. A generalized inverse of an $n \times m$ matrix A of any rank is an $m \times n$ matrix A^{g} such that for any vector Y for which $AX = Y$ is a consistent equation, $X = A^{g}Y$ is a solution. Penrose [43] shows that for any matrix A, there exists a unique matrix A^{g_p} satisfying the four conditions:

$$\begin{aligned} &\text{(i)} && AA^{g_p}A = A \\ &\text{(ii)} && A^{g_p}AA^{g_p} = A^{g_p} \\ &\text{(iii)} && (AA^{g_p})' = AA^{g_p} \\ &\text{(iv)} && (A^{g_p}A)' = A^{g_p}A . \end{aligned} \qquad (\dagger)$$

A^{g_p} is referred to as the Moore-Penrose inverse. However, a solution to $AX = Y$, where A is a singular square or rectangular matrix, requires a generalized inverse which satisfies only condition (i). Such a matrix is called a g_1-inverse of A and denoted by A^{g_1}. Likewise, the g_2- and g_3-inverse of A, denoted by A^{g_2} and A^{g_3}, are defined

respectively by the first two and the first three conditions of (†).

Before proceeding with the proof of Theorem 1.1.1, we state first the following lemma:

Lemma 1.1.1. A matrix G is a g_3-inverse of X iff it can be written as $G = (X'X)^{g_1}X'$.

Proof of Theorem 1.1.1. (Chipman [9]). Let $\hat{\beta} = My + d$. For $L'\hat{\beta}$ to be an unbiased estimator of $L'\beta$, we require

$$\begin{aligned} E(L'\hat{\beta}) &= E(L'My + L'd) \\ &= L'MX\beta + L'd \\ &= L'\beta , \end{aligned} \tag{3}$$

Condition (3) is satisfied iff

$$L'MX = L', \quad L'd = 0$$

or

$$a'XMX = a'X .$$

Thus for $M = X^{g_1}$, $L'My$ is an unbiased estimator of an estimable function $L'\beta$. Now

$$\begin{aligned} \text{var}\,(L'X^{g_1}y) &= \text{var}\,(a'XX^{g_1}y) \\ &= a'XX^{g_1}X^{g_1\prime}X'a\sigma^2 \\ &= a'XAA'X'a\sigma^2 \quad \text{where} \quad A = X^{g_1} . \end{aligned} \tag{4}$$

We wish to find an A which minimizes (4) subject to $XAX = X$.

Consider

$$XA[I - XX^{g_3}X^{g_3}{}'X']A'X' ,$$

since

$$I - XX^{g_3}X^{g_3}{}'X' = I - XX^{g_3}XX^{g_3} \quad \text{(condition (iii) of } g_3\text{-inverse)}$$

$$= I - XX^{g_3} \quad \text{(condition (ii) of } g_3\text{-inverse)}$$

$$= (I-XX^{g_3})(I-XX^{g_3}) \quad \text{(since idempotent)}$$

$$\geq 0 .$$

Thus $XAA'X'$ is minimal when A is a g_3-inverse of X. Since $(X'X)^{g_1}X' = X^{g_3}$ (Lemma 1.1.1) and any generalized inverse is a g_1-inverse, the proof is completed.

The two cases which have been discussed above may be treated within a unified framework. If $(X'X)$ is square and of full rank, $(X'X)^{g} = (X'X)^{-1}$. Moreover, as Rao [48] has pointed out, $(X'X)^{g}X'y$ and $\sigma^2(X'X)^{g}$ may be regarded as "estimates of β and the dispersion matrix of estimates respectively, for purposes of building up an estimate of any estimable function $L'\beta$ and determining its variance".

1.2 The Multicollinearity Problem

Linear relationships among the independent variables may exist in various forms, either between pairs of independent variables or in a more complicated manner involving several members of the independent set. In general, such intercorrelation results in:

(1) inaccurate estimation of parameters due to large sampling variances of the coefficients. As the columns of X become increasingly collinear,

the matrix (X'X) approaches singularity, resulting in the inverse matrix having some very large diagonal elements. In the limiting case, the determinant of (X'X) is zero and its inverse would not exist, leading to a completely indeterminate set of parameter estimates.

(2) uncertain specification of the model with respect to inclusion of variables and a danger that relevant variables will be discarded incorrectly. For example, if the i-th diagonal element is large, X_i may appear to be statistically insignificant even if it is important in the true relation.

(3) estimates of coefficients become very sensitive to slight changes in the data sample.

As a simple demonstration of the third difficulty, we consider the data in Table I concerning the imports, production, stock formation and consumption obtained from the French national accounts. These data reveal an approximate multicollinearity between production and consumption, namely $X_3 \simeq \frac{3}{4} X_1$.

Using least squares computer program MLREGR [42], we obtain,

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} -0.06788 \\ 0.58914 \\ 0.34725 \end{pmatrix}$$

TABLE I

Imports, Production, Stock Formation and Consumption in France (in Milliards of New Francs at 1956 Prices)

Year	y Imports	X_1 Gross Domestic Production	X_2 Stock Formation	X_3 Consumption
1949	15.9	149.3	4.2	108.1
1950	16.4	161.2	4.1	114.8
1951	19.0	171.5	3.1	123.2
1952	19.1	175.5	3.1	126.9
1953	18.8	180.8	1.1	132.1
1954	20.4	190.7	2.2	137.7
1955	22.7	202.1	2.1	146.0
1956	26.5	212.4	5.6	154.1
1957	28.1	226.1	5.0	162.3
1958	27.6	231.9	5.1	164.3
1959	26.3	239.0	0.7	167.6
1960	31.1	258.0	5.6	176.8
1961	33.3	269.8	3.9	186.6
1962	37.0	288.4	3.1	199.7
1963	43.3	304.5	4.6	213.9
1964	49.0	323.4	7.0	223.8
1965	50.3	336.8	1.2	232.0
1966	56.6	353.9	4.5	242.9

Source: E. Malinvaud, Statistical Methods of Econometrics (North-Holland: 1971).

and the estimated relation

$$y = -15.21577 - 0.06788X_1 + 0.58914X_2 + 0.34725X_3$$
$$\qquad\qquad\qquad\quad (-0.72478) \quad (2.96977) \quad (2.44402)$$

t values are given in parenthesis. The squared multiple correlation coefficient $R^2 = 0.9847$.

Suppose now the original model is re-estimated from data for 18 years, namely 1949 to 1966, instead of the original 15 years 1949 to 1963. A different set of parameter estimates is obtained and the estimated relation becomes

$$y = -19.73039 + 0.03210X_1 + 0.41421X_2 + 0.24293X_3$$
$$\qquad\qquad\qquad\quad (0.17198) \quad (1.28690) \quad (0.85253)$$

where the number in parenthesis refer again to t values . Here $R^2 = 0.9731$ and the sample correlations matrix is obtained as

$$C = \begin{pmatrix} 1.00000 & 0.21545 & 0.99893 \\ 0.21545 & 1.00000 & 0.21369 \\ 0.99893 & 0.21369 & 1.00000 \end{pmatrix} .$$

Table II gives the ratio of the original to revised estimates for the two parameters.

It can be seen that both $\hat{\beta}_2$ and $\hat{\beta}_3$ vary by more than 40%, while the coefficient $\hat{\beta}_1$ of X_1 in the revised model turns out to be positive. With the addition of three further years' data, the coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$, formerly significant at the 5% level ($t_{0.975,11} = 2.201$) are now insignificant ($t_{0.975,14} = 2.145$) . Thus

extension of the sample period has produced dramatic changes in the estimated relationship.

TABLE II

Ratio of Original to Revised Estimates

	$\left(\frac{\text{Original Estimate}}{\text{Revised Estimate}}\right) \times 100$
$\hat{\beta}_2$	142
$\hat{\beta}_3$	143

Difficulties at the computational level also arise in situations where multicollinearity is very severe, that is, when the determinant of $(X'X)$ is very close to but not zero. The observations of Klein and Nakamura [31] in this regard include the fact that while the elements of the inverse matrix $(X'X)^{-1}$ in the two variable case can still be calculated with sufficient accuracy by carrying enough digits at each computational stage, accurate estimation is considerably more difficult to attain when the matrix size is 5 by 5 or larger. Indeed, given the sort of intercorrelation frequently existing in econometric data, they note the virtual impossibility of calculating the inverses of 30 by 30 matrices, even if the most powerful electronic computer is available.

CHAPTER II

DETECTION OF MULTICOLLINEARITY

Two practical issues arise in connection with the problem of multicollinearity and its treatment. Firstly, how is its existence to be detected and its severity established? Secondly, how serious must multicollinearity be before it can be considered "harmful"? In attempting to answer these questions, several research workers have suggested various measures of multicollinearity and possible means of detection. The efforts of these workers are discussed below.

2.1 Measures of Multicollinearity

A common measure of the degree of multicollinearity is the value of the determinant of $\tilde{X}'\tilde{X}$, where $\tilde{X}$ is X normalized so that each column has zero mean and unit variance. $\tilde{X}'\tilde{X}$ is accordingly the sample correlations matrix C. The value of $|\tilde{X}'\tilde{X}|$ ranges from 0 to 1 as multicollinearity becomes less severe.

Another measure of multicollinearity is the von-Neumann and Goldstine [58] condition number

$$\lambda = \frac{\lambda_1}{\lambda_m}$$

where λ_1: the largest of the eigenvalues of $\tilde{X}'\tilde{X}$

λ_m: the smallest of the eigenvalues of $\tilde{X}'\tilde{X}$.

If the columns of $\tilde{X}$ are orthogonal, λ will be 1. However, as they become collinear, λ_m will become very small, so that λ becomes quite large. As λ increases, the probability of significant error of estimates also increases.

Between these two measures, λ is to be preferred owing to its more direct relationship with the effect of multicollinearity. Notwithstanding this, disadvantages exist in that both measures do not give information about the pattern of interdependence, and that in neither of them can absolute conceptions of bigness or smallness be fixed. Nevertheless, in the case of λ, some guidelines may be obtained from the fact that for a correlation matrix

$$C = (c_{ij})$$

which has been scaled according to

$$\frac{0.99}{4} \leq \max_{j=1,\ldots,m \text{ or } i=1,\ldots,m} \left(\sum_{i=1}^{m} c_{ij} \times c_{ij}, \sum_{j=1}^{m} c_{ij} \times c_{ij} \right) \leq 0.99$$

λ has a minimum value of 1 and is $\sim m$ with probability asymptotically 1.

Multicollinearity can be localized by calculation of the determinant of $(\tilde{X}'\tilde{X})$ and those of $(\tilde{X}'\tilde{X})_{ii}$, which are matrices obtained by omitting each of the independent variables in turn. If $\tilde{X}_i$ is orthogonal to the other members of $\tilde{X}$, then

$$|(\tilde{X}'\tilde{X})_{ii}| = |\tilde{X}'\tilde{X}|$$

and the diagonal element c^{ii} of $(\tilde{X}'\tilde{X})^{-1}$, because it is equal to

$$c^{ii} = \frac{|(\tilde{X}'\tilde{X})_{ii}|}{|\tilde{X}'\tilde{X}|} ,$$

would have the value 1 . This can be shown to be its minimum value in the following way. Defining $R^2_{X_i}$ as the squared multiple correlation between X_i and the other members of X , we have

$$c^{ii} = \frac{1}{1 - R^2_{X_i}}$$

and

$$0 \leq R^2_{X_i} \leq 1 .$$

Thus

$$|(\tilde{X}'\tilde{X})_{ii}| \geq |\tilde{X}'\tilde{X}| .$$

When $\tilde{X}_i$ is perfectly dependent on the remaining members of $\tilde{X}$, $|\tilde{X}'\tilde{X}|$ vanishes while $|(\tilde{X}'\tilde{X})_{ii}|$, since it does not contain $\tilde{X}_i$, remains unaffected. In this case, $c^{ii} = \infty$. If perfect linear dependency exists in $(\tilde{X}'\tilde{X})_{ii}$, then $c^{ii} = 0/0$ which is indeterminate. However, one would not have proceeded to localize multicollinearity if the determinant is found to be exactly 0 in the first place. Therefore, the size of the diagonal elements of $(\tilde{X}'\tilde{X})^{-1}$, $1 \leq c^{ii} \leq \infty$, is a good indicator of the location of the problem.

To decide whether multicollinearity is harmful, a number of rules-of-thumb have been proposed. Farrar and Glauber [15] suggest the rule

$$\max_i (R^2_{X_i}) < R^2 .$$

In other words, the overall R^2 should exceed the highest R^2 of any regression of one independent variable on its couterparts. In a recent review, Raduchel [46] has explicated this rule in terms of the generalized variance of the coefficients, that is, the determinant of their variance-covariance matrix. Defining ρ_i as the ratio of generalized variance of the coefficients of a regression including X_i to the generalized variance of the coefficients of a regression excluding it, and applying standard theorems on correlation coefficients, he obtains

$$\rho_i = \frac{(1-r_i^2)^{m-1}(1-R^2)}{(1-R_{X_i}^2)}$$

where r_i is the partial correlation of y and X_i, given the influence of the remaining independent variables.

Farrar-Glauber's rule of thumb therefore guarantees that all ρ_i will be less than 1. As a modification of this rule, Haitovsky [22] has suggested that comparison should be made instead between the partial correlation coefficients of all pairs of the independent variables and the overall R^2. His views will be published in a forthcoming paper [23].

Turning now to the subject of detection, a comprehensive search of the literature revealed as many as four methods have been proposed since 1934. The earliest attempt to deal with the problem goes back to Frisch [18].

2.2 Frisch Bunch Map Analysis

Essentially the basic idea of this technique is the determination of the regression plane by minimizing the residual sums of squares in various directions, for each of all possible subsets as well as the full set of variables including the dependent variable y (usually denoted by X_1).

The construction of bunch maps for any subset $X_1, X_2, \ldots, X_k$, $2 \leq k < m+1$, or the full set involves only two variables in each map. For example, a subset consisting of three variables will have ${}_3C_2$ bunch maps. Each bunch map consists of k beams, the k-th beam having slope $-R_{kj}/R_{ki}$, where $i < j$ and R_{kj} denotes the cofactor of r_{kj} in the correlation matrix C.

Using standardized variables, it can easily be shown that $-R_{kj}/R_{ki}$ is simply the ratio of the coefficients of X_j and X_i in the regression of X_k on $X_1, X_2, \ldots, X_{k-1}$.

Following construction of the bunch maps, each bunch is compared with the corresponding bunch in the first subsets of the set considered, comparison being in terms of the dispersion of the beams or their lengths. When the inclusion of a variable renders the new bunch more widely spread, the variable is deduced to be correlated with the other variables in the bunch. Conversely, if the variable added possess a very short beam relative to the other beams, it is orthogonal to the other variables. A theoretical explanation of these deductions has been given by Malinvaud [36].

An elucidation of the use of Frisch's Bunch Map analysis is provided by considering the bunch maps (Figure 1) which have been constructed from the data in Table III. Focussing first on the bunch (24), we see that the bunch remains more or less unchanged upon the inclusion of variable No. 3. In addition, the beam corresponding to variable No. 3 is extremely short. We therefore conclude that variable No. 3 is approximately orthogonal to the other variables. Multicollinearity on the other hand is exemplified by the behaviour of the bunch (123) when variable No. 4 is added to it. Since the bunch becomes less tight, we deduce that variable No. 4 is correlated with the other variables in the set.

To a certain extent then, Frisch's technique involves subjectivity in interpretation of the bunch maps. This lack of precision, as well as the laborious calculation required for all the cofactors have rendered the technique obsolete. Another early technique that has been developed but which fared no better, was that by Tintner [54].

2.3 Tintner Method

Tintner adopts Frisch's view that the variables are composed of two parts,

$$X_{it} = M_{it} + y_{it}, \quad \begin{matrix} i = 1,\ldots,m \\ t = 1,\ldots,n \end{matrix}$$

where M_{it} is the systematic or "true" part and y_{it} is the random or "error" component which arises as error of measurement. The y_{it} are supposed to be normally distributed with mean zero.

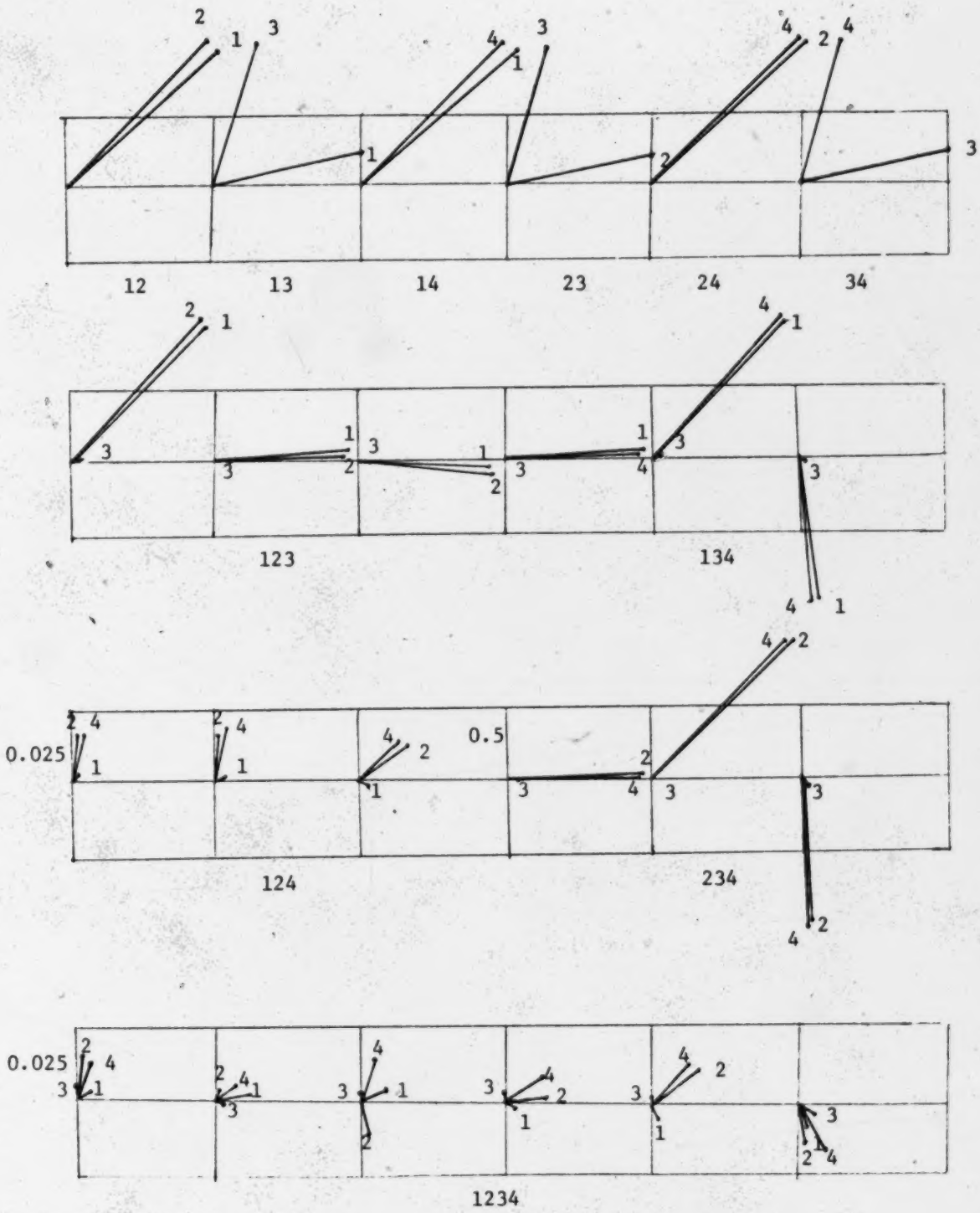

FIGURE 1

Bunch Maps for Variables in the French Economy Data

TABLE III

Adjoints for Subsets and Full Set of Variables in the French Economy Data

R_{ij}	1	2
1	1.000000	-0.984180
2		1.000000

R_{ij}	1	3
1	1.000000	-0.265910
		1.000000

R_{ij}	1	4
1	1.000000	-0.984770
4		1.000000

R_{ij}	2	3
2	1.000000	-0.215450
3		1.000000

R_{ij}	2	4
2	1.000000	-0.998930
4		1.000000

R_{ij}	3	4
3	1.000000	-0.213690
4		1.000000

R_{ij}	1	2	3
1	0.953580	-0.926890	-0.053870
2		0.929290	0.046250
3			0.031389

R_{ij}	1	2	4
1	0.002137	-0.000463	-0.001642
2		0.030209	-0.029721
4			0.030370

TABLE III

R_{ij}	1	3	4
1	0.954320	-0.055470	-0.927940
3		0.030230	0.048170
4			0.929280

R_{ij}	2	3	4
2	0.954350	-0.001989	-0.952900
3		0.002139	0.001529
4			0.953450

R_{ij}	1	2	3	4
1	0.002037	-0.000332	-0.000055	-0.001649
2		0.025770	-0.000034	-0.025409
3			0.000064	0.000137
4				0.027032

We assume that the variance-covariance matrix $[c_{ij}]$ of the errors is known or that it can be estimated, for example, by the Variate Difference Method. Let $[V_{ij}]$ denote the estimate of $[c_{ij}]$. Then, to estimate the number of independent linear relationships among the M_{it} in the hypothetically infinite population which corresponds to our sample, Tintner suggests that the following determinantal equation be formed

$$\begin{vmatrix} a_{11}-\lambda V_{11} & a_{12}-\lambda V_{12} & \cdots & a_{1m}-\lambda V_{1m} \\ a_{21}-\lambda V_{21} & a_{22}-\lambda V_{22} & \cdots & a_{2m}-\lambda V_{2m} \\ & \cdots\cdots\cdots & & \\ a_{m1}-\lambda V_{m1} & a_{m2}-\lambda V_{m2} & & a_{mm}-\lambda V_{mm} \end{vmatrix} = 0 \qquad (5)$$

where a_{ij} is the covariance of X_i and X_j and a_{ii} the variance of X_i.

We form the test function

$$\Lambda_r = (n-1)(\lambda_1+\lambda_2+\ldots+\lambda_r)$$

where λ_1 is the smallest root of (5), λ_2 is the next smallest and so on. According to Hsu [27], Λ_r is asymptotically χ^2 distributed with $r(n-m-1+r)$ degrees of freedom. In addition, Anderson [1] has shown that

$$\frac{\Lambda_r - nr}{\sqrt{2nr}}$$

has an asymptotic normal distribution with mean 0 and variance 1.

Therefore, if $\Lambda_1, \Lambda_2, \ldots, \Lambda_q$ are not significant, but Λ_{q+1} is, we can conclude that there are q independent linear relationships among the M_{it}.

Tintner published his technique in 1952 but since then it has seen little use due to a number of inherent weaknesses. These include the facts that it is valid only for large samples and that its appropriateness is conditional upon the existence of error of observations. Moreover, it relies on the assumption that the variance-covariance matrix of the errors can be known or estimated.

2.4 Farrar-Glauber Technique

In recent years, more satisfactory methods have emerged, one major contribution being that of Farrar and Glauber [15]. Viewing multicollinearity as a feature of the sample rather than that of the population, they have defined multicollinearity in terms of departures from orthogonality. Under the assumption that the sample is taken from an orthogonal, multivariate normal population, they then propose a three level test for the "presence, location and severity of multicollinearity". At the primary level of detection, the determinant of $(\tilde{X}'\tilde{X})$ is transformed into

$$\chi^2_{|\tilde{X}'\tilde{X}|}(v) = -[n - 1 - \tfrac{1}{6}(2m+5)]\log |\tilde{X}'\tilde{X}|$$

which was shown by Bartlett [2] to have an approximate Chi square distribution with $v = m(m-1)/2$ degrees of freedom under the null hypothesis that the columns of X are orthogonal ($|\tilde{X}'\tilde{X}| = 1$). Next, to determine

which particular variable is affected by multicollinearity, diagonal elements of $(\tilde{X}'\tilde{X})^{-1}$ are transformed in such a way as to enable the use of F test. By applying the results obtained by Wilks [60] on the distribution of the ratio of the determinant of correlation matrix to h of its principal minors which are mutually exclusive, and making the transformation

$$w = (c^{ii}-1)\left(\frac{v_1}{v_2}\right) \tag{6}$$

where $v_1 = n - m$, $v_2 = m - 1$. Farrar and Glauber derive the density function of w as an F distribution with v_1 and v_2 degrees of freedom.

Since $c^{ii} = \frac{1}{1 - R^2_{X_i}}$, (6) can be written as

$$w = \left(\frac{R^2_{X_i}}{1-R^2_{X_i}}\right)\left(\frac{v_1}{v_2}\right)$$

Finally, a notion of the pattern of interdependence can be obtained by examining the partial correlation coefficients of the variables. Farrar and Glauber show that normalized off - diagonal elements of $(\tilde{X}'\tilde{X})^{-1}$ yield the partial correlation coefficients among the independent variables, namely, for any pair $\tilde{X}_i$, $\tilde{X}_j$,

$$c_{ij\cdot} = \frac{-c^{ij}}{\sqrt{c^{ii}\cdot c^{jj}}} .$$

The t test is used as a criterion since the statistic

$$t_{ij\cdot}(v) = \frac{c_{ij\cdot}\sqrt{n-m}}{\sqrt{1-c_{ij\cdot}^2}}$$

has a t distribution with $v = n - m$ degrees of freedom.

Nevertheless, as Haitovsky has pointed out, since economic data are hardly ever orthogonal, this test is of little meaning to him. He suggests an alternative method using the statistic

$$\chi_H^2(v) = -[-n - 1 - \tfrac{1}{6}(2m+5)] \log (1-|\tilde{X}'\tilde{X}|) .$$

In this context, the null hypothesis becomes $|\tilde{X}'\tilde{X}| = 0$, that is, the data are perfectly collinear. The value of χ_H^2 would be small when multicollinearity is high since $|\tilde{X}'\tilde{X}|$ would approach zero. The severity of multicollinearity can be measured by the level of significance at which the null hypothesis is accepted. A χ^2 value, for example, significant at the 0.9 level, would indicate a high degree of multicollinearity.

In a recent review, Raduchel expresses agreement with Haitovsky's comment on the test of Farrar and Glauber. He criticizes though the usefulness of Haitovsky's proposal of a heuristically motivated test of the converse hypothesis. The situation of perfect collinearity is just as unlikely as the other extreme in practice, or when it does, there would be little meaning in applying regression anyway.

2.5 Kirchdorfer Procedure

The latest development in the problem of detection comes from the German statistician Kirchdorfer [30]. His method is based on the Gram-Schmidt orthogonalization process. Suppose X is factored into an orthonormal matrix D and an upper triangular matrix U.

$$X = DU . \tag{7}$$

The elements of D and U are determined by a process involving an intermediate matrix E. Let x_{ij} denote the elements of X, with x_{i0}, the elements of the variable X_0, equal to 1 for all i, $i = 1,\ldots,n$. The elements of D, U, and E, denoted by d_{ij}, u_{ij}, and e_{ij} respectively, are obtained in the following way:
Beginning with $k = 0$, let

$$e_{ik} = x_{ik} \quad \text{for} \quad i = 1,2,\ldots,n \tag{8}$$

$$u_{kk} = \sqrt{\sum_{i=1}^{n} e_{ik}^2} \tag{9}$$

$$d_{ik} = \frac{e_{ik}}{u_{kk}} \quad \text{with} \quad i = 1,2,\ldots,n \tag{10}$$

$$u_{kj} = \sum_{i=1}^{n} d_{ik} x_{ij} \quad \text{with} \quad j = k+1,k+2,\ldots,m \tag{11}$$

and then with $k = k + 1$, let

$$e_{ik} = x_{ik} - \sum_{j=0}^{k-1} u_{jk} d_{ij} \quad \text{for} \quad i = 1,2,\ldots,n . \tag{12}$$

The above procedure is repeated from (9) to (12) until $k = m$.

By simple manipulation of the original model, Kirchdorfer derives the result

$$\hat{\beta} = U^{-1}D'y \ .$$

Since

$$\begin{aligned}(X'X)^{-1} &= (U'D'DU)^{-1} \\ &= (U'U)^{-1} \quad \text{(by orthonormality of D)} \\ &= U^{-1}U'^{-1} \end{aligned} \tag{13}$$

the process of detecting multicollinearity is much simplified by considering the matrix U instead. An examination of the size of the diagonal elements of U would indicate the source of multicollinearity. In the general case, if u_{ii} is small, Kirchdorfer concludes that X_i is correlated with the remaining independent variables.

Kirchdorfer's procedure is a recently invented one and since its publication in 1971, no reports have appeared in the literature concerning either its theory or practice. It therefore seemed of interest to apply the technique to our data on the French economy (Table I). Using Program A (Appendix II), the resultant matrix U is obtained as follows:

$$U = \begin{pmatrix} 4.24264 & 1007.69751 & 15.60349 & 710.12354 \\ 0 & 261.88599 & 1.54679 & 171.25647 \\ 0 & 0 & 7.01129 & -0.26582 \\ 0 & 0 & 0 & 7.91342 \end{pmatrix} .$$

Examining the diagonal elements, it is seen that none of them have small value. More importantly, variable No. 4 is not identifiable on the basis of Kirchdorfer's procedure as the source of multicollinearity. Yet as we recall, the calculations of Chapter I and the Bunch Map analysis performed earlier in this chapter, point to the existence of multicollinearity between the variables No. 2 and 4. Thus our particular set of data constitutes a counterexample to Kirchdorfer's procedure.

How may this phenomenon be explained? Consider again equation (13)

$$(X'X)^{-1} = U^{-1}U'^{-1}$$

where U is an upper triangular matrix. If we let x^{ii} be the diagonal elements of $(X'X)^{-1}$, then

$$x^{ii} = \frac{1}{u_{11}^2} + (u^{12})^2 + (u^{13})^2 + \ldots + (u^{1m})^2$$

$$x^{22} = \frac{1}{u_{22}^2} + (u^{23})^2 + (u^{24})^2 + \ldots + (u^{2m})^2$$

$$\ldots\ldots\ldots$$

$$x^{(m-1)(m-1)} = \frac{1}{u_{(m-1)(m-1)}^2} + (u^{(m-1)(m)})^2$$

$$x^{mm} = \frac{1}{u_{mm}^2} \qquad (14)$$

where u^{ij} is the ij-th element in the matrix U^{-1}.

We know that when X_i is multicollinear with other members of the independent set, x^{ii} will be large. From our set of equations in (14), it is clear that u_{ii} will likely be very small when x^{ii} is large only in the case $i = m$. For all $i = 1,2,\ldots,m-1$, the presence of additional squared terms in the set of equations (14) may result in $1/u_{ii}$ not being necessarily large even if x^{ii} is large.

In sum, Kirchdorfer's method succeeds with certainty only for the case when the m-th variable is affected by multicollinearity. The diagonal elements of U in all other instances of multicollinearity may or may not be small, so that no certain indication is obtained as to whether or not the problem exists.

Kirchdorfer's own numerical example illustrates the above argument well. It is coincidental that only the third variable in his set of X_i's is affected by multicollinearity, as evidenced by the significantly small value of u_{33}. If multicollinearity had been located in other variables, and not in the last variable X_3, then u_{ii} may not be necessarily small.

CHAPTER III

RESOLVING MULTICOLLINEARITY

The discussion thus far has delineated multicollinearity in terms of its detection by a variety of methods. Given that the problem indeed exists in a particular situation, the next logical task to face is the search for meaningful remedy. A number of alternative means of resolution have been proposed, some for cases where X is less than full rank, and others applying to cases where multicollinearity is only approximate. The latter category of procedures are less easily accomplished than those used to resolve situations of perfect collinearity on account of the need for a priori information or tedious computation. What follows is a review of the procedures of both categories that have been suggested to date.

3.1 Generalized Inverses

One approach to the estimation of the linear model of less than full rank using generalized inverses has been discussed in Chapter I. In an alternative approach, the parameters are subjected to linear constraints of the form

$$R\beta = c$$

where R is an $n \times (m-r)$ matrix of rank s and $s \leq (m-r) < n$.

Minimization of the constrained sum of squares leads to

equations of the form

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix}\begin{bmatrix} \tilde{\beta} \\ \gamma \end{bmatrix} = \begin{bmatrix} X'y \\ c \end{bmatrix}$$

or

$$M\begin{bmatrix} \tilde{\beta} \\ \gamma \end{bmatrix} = \begin{bmatrix} X'y \\ c \end{bmatrix} \tag{15}$$

where γ is a vector of Lagrange's multipliers. We have the following lemma

Lemma 3.1.1. The g_1-inverse of M is given by

$$M^{g_1} = \begin{pmatrix} K^{g_1} - K^{g_1}R'G^{g_1}RK^{g_1} & K^{g_1}R'G^{g_1} \\ G^{g_1}RK^{g_1} & G^{g_1}G-G^{g_1} \end{pmatrix} \tag{16}$$

where $K = (X'X+R'R)$ and $G = RK^{g_1}R'$.

The derivation of Lemma 3.1.1 is based on the following results due to Bose [6] and Rao [47].

(1) (i) $X(X'X)^{g_1}X'X = X$

(ii) $X'X(X'X)^{g_1}X' = X'$

(iii) $AX^{g_1}X = A$ iff A is contained in the row space of X .

(2) If H is a $q \times k$ matrix such that the row space of H is contained in the row space of the $k \times k$ matrix $S = (X'X)$, then

(i) $r(HS^{g_1}H') = r(H)$ where $r(H)$ denotes the rank of H .

(ii) $HS^{g_1}H'$ is unique and positive definite.

Proof of Lemma 3.1.1. (Dwyer [13]). Let M be premultiplied by F, where

$$F = \begin{pmatrix} I & R' \\ 0 & I \end{pmatrix}$$

so that

$$FM = \begin{pmatrix} s+R'R & R' \\ R & 0 \end{pmatrix}.$$

Now

$$S + R'R = K$$

$$RK^{g_1}R' = G$$

and consider the reduction $B(FM)Q$, where

$$B = \begin{pmatrix} I & 0 \\ -RK^{g_1} & I \end{pmatrix}$$

and

$$Q = \begin{pmatrix} I & -K^{g_1}R' \\ 0 & I \end{pmatrix}.$$

Let $P = BF$, then P is clearly nonsingular. Furthermore,

$$K = [X' \quad R'][\begin{smallmatrix} X \\ R \end{smallmatrix}]$$

$$= T'T \quad \text{(say)}.$$

By (i) and (ii) of 1,

$$T(T'T)^{g_1}T'T = T$$

and

$$T'T(T'T)^{g_1}T' = T'$$

thus the following relations hold

$$RK^{g_1}K = R \tag{17}$$

$$KK^{g_1}R' = R' \tag{18}$$

$$XK^{g_1}K = X \tag{19}$$

$$KK^{g_1}X' = X' \ . \tag{20}$$

It follows from (17), (18) and (19) that

$$PMQ = \begin{pmatrix} K & 0 \\ 0 & -G \end{pmatrix}$$

and

$$M^{g_1} = Q\begin{pmatrix} K^{g_1} & 0 \\ 0 & -G^{g_1} \end{pmatrix}P$$

with

$$P = \begin{pmatrix} I & R' \\ -RK^{g_1} & I-G \end{pmatrix} .$$

By (17) and (i) and (ii) of 2 , $G = RK^{g_1}R'$ is unique, positive definite and has the same rank as R . It then follows from 1 (iii) that, since the row space of R' and G are the same,

$$R'G^{g_1}G = R' \ .$$

On completion of the reduction, M^{g_1} is obtained as given in (16).

It is evident that a solution to equation (15) is given by

$$\tilde{\beta} = (K^{g_1} - K^{g_1}R'G^{g_1}RK^{g_1})X'y + K^{g_1}R'G^{g_1}c \ . \qquad (21)$$

Rao [47] has shown that $L'\tilde{\beta}$, with $\tilde{\beta}$ as given in (21), is the best linear unbiased conditional estimate of an estimable function $L'\beta$.

Plackett [45] considers the case where the restrictions are chosen in such a way that $[X' \ R']$ has the full rank m (R is thus complementary to X) and obtains the minimum variance conditionally unbiased estimator of β as

$$\tilde{\beta} = (X'X + R'R)^{-1}(X'y + R'c) \ .$$

A further illustration of Plackett's solution in terms of g_3-inverses is provided by Chipman [9]. Defining

$$Z = \begin{bmatrix} X \\ R \end{bmatrix}$$

we have

$$\begin{aligned} Z^{g_3} &= (Z'Z)^{-1}Z' \\ &= (X'X+R'R)^{-1}[X' \ R'] \\ &= [(X'X+R'R)^{-1}X' \quad (X'X+R'R)^{-1}R'] \\ &= I \ . \end{aligned} \qquad (22)$$

Since X and R are complementary,

$$\begin{aligned} X(X'X+R'R)^{-1}R' &= (R(X'X+R'R)^{-1}X')' \\ &= 0 \ . \end{aligned}$$

Thus premultiplying (22) by X and postmultiplying by $(X'X+R'R)^{-1}X'$, conditions (i) and (ii) of g_3-inverse are verified respectively. Verification of condition (iii) is trivial. Performing similar operations on $(X'X+R'R)^{-1}R'$, we thus have

$$\tilde{\beta} = X^{g_3}y + R^{g_3}c .$$

3.2 Using Additional Data

One proposed remedy involves the acquisition of additional data, which however may not always be available. In the fortunate case when the researcher has access to new data, an efficient criterion of selection which would best reduce the standard error of estimates has been suggested by Silvey [51]. He points out that for the parametric function $L'\beta$ to be estimable, L must be a linear combination of eigenvectors of $(X'X)$ corresponding to non-zero roots. L can thus be written in the form

$$L = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_j v_j$$

where v_i are normalized eigenvectors of $(X'X)$. Utilizing the fact that

$$\text{var}\,(L'\hat{\beta}) = \sigma^2\left(\frac{\alpha_1^2}{\lambda_1} + \frac{\alpha_2^2}{\lambda_2} + \ldots + \frac{\alpha_j^2}{\lambda_j}\right)$$

which implies that the bigger any λ_i, the smaller is its contribution to the variance, Silvey is thus able to show that precise estimation is possible in the direction of eigenvectors corresponding to large eigenvalues.

Silvey's selection criterion is to assume an additional observation y_{n+1} is taken at the values $x'_{n+1} = (x_{1,n+1}, x_{2,n+1}, \ldots, x_{m,n+1})$ of the independent variables, where $x'_{n+1} = \ell v_i$, ℓ some non-zero number.

The new model is

$$\begin{pmatrix} y \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} X \\ x'_{n+1} \end{pmatrix} \beta + \begin{pmatrix} u \\ u_{n+1} \end{pmatrix}$$

or

$$y_* = X_* \beta + u_* .$$

Then

$$\begin{aligned} X'_* X_* &= X'X + x_{n+1} x'_{n+1} \\ &= X'X + \ell^2 v_i v'_i \end{aligned}$$

$$\begin{aligned} X'_* X_* v_i &= X'X v_i + \ell^2 v_i v'_i v_i \\ &= (\lambda_i + \ell^2) v_i . \end{aligned}$$

Thus v_i is an eigenvector of $X'_* X_*$ corresponding to the root $(\lambda_i + \ell^2)$. In addition,

$$\begin{aligned} X'_* X_* v_j &= X'X v_j + \ell^2 v_i v'_i v_j \\ &= \lambda_j v_j \quad \text{since } v'_i v_j = 0 \end{aligned}$$

so that v_j is an eigenvector of $X'_* X_*$ corresponding to root λ_j. The eigenvectors $(X'X)$ are thus those of $X'_* X_*$ and all eigenvalue are the same except that λ_i is now increased to $(\lambda_i + \ell^2)$. Therefore,

if the new independent variables are chosen in the direction of eigenvectors of $(X'X)$ corresponding to small roots, the standard error would be reduced.

In the case when the new observation x_{n+1} is not necessarily in the direction of an eigenvector of $(X'X)$, Silvey shows that the optimum direction of x_{n+1}, subject to the condition $x'_{n+1}x_{n+1} = b^2$ is that of the vector $(I+b^{-2}X'X)^{-1}L$, which holds for both singular or nonsingular X. This same result has been obtained by Gupta [21] in a simpler and more concise fashion.

Researchers have also grappled with the problem of treating multicollinearity when additional data are not readily available. Of several possible remedies that have appeared, some have been more successful than others. The methods presently available will now be discussed.

3.3 Incorporating Extraneous Information

One procedure which has seen extensive use involves the incorporating of information extraneous to the sample followed by re-estimation of the regression equation. An investigator, for instance, may have knowledge of the ratio of some coefficients. Alternatively, the values of certain coefficients or their linear combinations may be known. This procedure of using extraneous information varies in form according to the type of information available.

The method commonly employed by econometricians involves

combining time-series and cross-sectional samples. An example is seen in demand studies where income and prices in time-series data are usually collinear. Cross-sectional samples are then used for estimating the income coefficients. The procedure may be formulated in the following way.

Suppose that we have estimates of $(m-p)$ of the m elements of β. Without loss of generality, we may renumber the X variables so that the estimated coefficients refer to the last $(m-p)$ variables. The coefficients of the first p variables are then to be estimated. Consider the partitioned relationship

$$y = X_p\beta_p + X_{m-p}\beta_{m-p} + u . \tag{23}$$

Let $\tilde{\beta}^*_{m-p}$ be an unbiased estimator of β_{m-p}

$$\tilde{\beta}^*_{m-p} = \beta_{m-p} + d \tag{24}$$

with $Ed = 0$ and the variance-covariance matrix is known, $E(\tilde{\beta}^*_{m-p}-\beta_{m-p})(\tilde{\beta}^*_{m-p}-\beta_{m-p})' = V^*_{m-p}$, and assume $Edu' = 0$. We regress $y^* = y - X_{m-p}\tilde{\beta}^*_{m-p}$ on $X_p\beta_p$ to estimate β_p as

$$\tilde{\beta}^*_p = (X'_pX_p)^{-1}X'_py^* = (X'_pX_p)^{-1}X'_p(y-X_{m-p}\tilde{\beta}^*_{m-p}) .$$

Substitution of (23) gives

$$\tilde{\beta}^*_p = \beta_p + (X'_pX_p)^{-1}X'_pu - (X'_pX_p)^{-1}X'_pX_{m-p}(\tilde{\beta}^*_{m-p}-\beta_{m-p}) .$$

Since

$$E(u) = 0$$

and

$$E(\tilde{\beta}^*_{m-p}) = \beta_{m-p}$$

thus

$$E(\tilde{\beta}^*_{p}) = \beta_{p}$$

and from the assumption of independence of the two sets of data, the variance-covariance matrix for $\tilde{\beta}^*_p$ is

$$V^*_p = E(\tilde{\beta}^*_p - \beta_p)(\tilde{\beta}^*_p - \beta_p)'$$

$$= \sigma^2 (X'_p X_p)^{-1} + (X'_p X_p)^{-1} X'_p X_{m-p} V^*_{m-p} X'_{m-p} X_p (X'_p X_p)^{-1}$$

σ^2 can be estimated by $\hat{u}'\hat{u}/n-m$ where

$$\hat{u} = y^* - X_p \tilde{\beta}^*_p = y - X_p \tilde{\beta}^*_p - X_{m-p} \tilde{\beta}^*_{m-p} .$$

One shortcoming of the above procedure has been attributed to the fact that cross-sectional data are by nature usually long-run, whereas annual time-series data are often short-run in character. As Kuh and Meyer [33] has pointed out, the combination of different structures to overcome multicollinearity is improper, and leads in fact to discrepancies in the estimation.

As the second variant of the procedure of using extraneous information, suppose that the extreneous information consists of exact linear restrictions on the coefficients,

$$H'\beta = h$$

where h is a $k \times 1$ known vector and H' an $k \times m$ known matrix of rank $k < m$. We have, therefore, k independent restrictions on the elements of β.

In incorporating this information, we utilize the method of Lagrange's multipliers to determine the estimator of β which minimize $(y-X\beta)'(y-X\beta)$ subject to the restriction $H'\beta - h = 0$. It can easily be shown that the solution for the coefficient estimator under constraint is

$$\tilde{\beta}^* = \hat{\beta} + (X'X)^{-1}H[H'(X'X)^{-1}H]^{-1}(h-H'\hat{\beta}) . \qquad (25)$$

Substitution of $\hat{\beta} = \beta + (X'X)^{-1}X'u$ into (25) yields

$$\tilde{\beta}^* = \beta + (X'X)^{-1}X'u + (X'X)^{-1}H[H'(X'X)^{-1}H]^{-1}[h-H'\beta-H'(X'X)^{-1}X'u] .$$

Therefore, if $H'\beta = h$ is true, $\tilde{\beta}^*$ would be unbiased. The variance-covariance matrix of $\tilde{\beta}^*$ is given by

$$V^* = \sum_{\tilde{\beta}^*\tilde{\beta}^*} = V - VH(H'VH)^{-1}H'V$$

where $V = \sum_{\hat{\beta}\hat{\beta}} = \sigma^2(X'X)^{-1}$ is the variance-covariance matrix of the ordinary least squares estimator $\hat{\beta}$.

This estimator $\tilde{\beta}^*$ has been shown by Theil [52] to be the best linear unbiased estimator of β in the class of all unbiased estimators which are linear functions of y and h, provided $H'\beta = h$ is true.

The third context involving use of extraneous information is a method which utilizes both the extraneous and sample information to

estimate β_{m-p} and β_p efficiently. We assume that the extraneous information may be represented by

$$r = H'\beta + d \ . \tag{26}$$

Combining (26) with the basic model, we have

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ H' \end{pmatrix}\beta + \begin{pmatrix} u \\ d \end{pmatrix} .$$

The variance-covariance matrix of the extended disturbance is

$$E \begin{pmatrix} u \\ d \end{pmatrix}[u \quad d]' = \begin{pmatrix} \sigma^2 I & 0 \\ 0 & \psi \end{pmatrix} \tag{27}$$

where

$$\psi = Edd'.$$

The zero nature of the off-diagonal submatrices results from the assumption of independence between the sample and prior information. Since the variance-covariance matrix (27) is not $\sigma^2 I$, ordinary least squares cannot be used. An application of Aitken's generalized least squares procedure leads to estimates

$$\tilde{\beta}^* = \left([X' \quad H] \begin{pmatrix} \sigma^2 I & 0 \\ 0 & \psi \end{pmatrix}^{-1} \begin{pmatrix} X \\ H' \end{pmatrix} \right)^{-1} [X' \quad H] \begin{pmatrix} \sigma^2 I & 0 \\ 0 & \psi \end{pmatrix}^{-1} \begin{pmatrix} y \\ r \end{pmatrix}$$

$$= \left(\frac{1}{\sigma^2} X'X + H\psi^{-1}H'\right)^{-1} \left(\frac{1}{\sigma^2} X'y + H\psi^{-1}r\right) .$$

Specifically, for the example in the previous section, where

we have estimates of the coefficients of the last (m-p) variables

$$r = \tilde{\beta}^{*}_{m-p} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\beta}^{*}_{p+1} \\ \vdots \\ \tilde{\beta}^{*}_{m} \end{pmatrix}$$

and

$$H' = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & \\ \cdot & 1 & 0 \cdots & 0 \\ \cdot & 0 & \ddots & \vdots \\ \cdot & \vdots & 1 & \\ \cdot & \vdots & & \ddots\ 0 \\ 0 & 0 & \cdots & 0\ 1 \end{pmatrix} .$$

$\tilde{\beta}^{*}$ has the property of being the best linear unbiased estimator "best" refers to both extraneous and sample information. The variance-covariance matrix of this estimator is given by

$$\text{var}\ (\tilde{\beta}^{*}) = [\frac{1}{\sigma^2} X'X + H\psi^{-1}H']^{-1} .$$

A shortcoming of this method lies in the fact that knowledge of σ^2 and ψ is required. This difficulty can be circumvented by employing unbiased estimators of these variances and covariances. Theil [53], in search of a heuristic procedure, has suggested the following conditional estimator of β

$$\tilde{\beta}^{*} = (\frac{1}{s^2} X'X + H\psi^{-1}H')^{-1}(\frac{1}{s^2} X'y + H\psi^{-1}r)$$

where

$$s^2 = y'[I - X(X'X)^{-1}X']^{-1}X'y/n-m \ .$$

Alternatively, the investigator may have knowledge about the bounds on the values of some coefficients. Suppose it is known a priori that the coefficient β_1 lies between 0 and 1, probably between $\frac{1}{4}$ and $\frac{3}{4}$. This knowledge can be formulated as

$$\frac{1}{2} = \beta_1 + d$$

with

$$Ed = 0$$

$$Ed^2 = \frac{1}{16}$$

so that $\beta_1 \pm \sigma_{\beta_1}$ gives a range from $\frac{1}{4}$ to $\frac{3}{4}$ and $\beta_1 \pm 2\sigma_{\beta_1}$ gives a range from 0 to 1. The procedure described in the preceding paragraph is then applicable to obtain the best linear estimator. In this case, we set

$$r = \frac{1}{2}$$

$$H' = [1\ 0\ \ldots\ldots\ 0]$$

and

$$\psi = \frac{1}{16} \ .$$

Knowledge about linear combinations of coefficients may also be handled by a similar procedure to the above.

Finally, we may have situations in which the extraneous information take on the form

$$H'\beta \leq h .$$

In this event, the familiar technique of quadratic programming is then applied to minimize the sum of squared deviations $(y-X\beta)'(y-X\beta)$ subject to the constraints.

We note in conclusion that the utilization of extraneous information, as a procedure of resolving multicollinearity, leads to better estimates. One shortcoming though may be said to reside in the fact that a priori information often consists of facts or relationships adduced from expirical economic studies. The validity of this information is therefore always a problem to be faced. Fox [17] has in fact quite strongly stated that "if we use purely arbitrary coefficients to get around a statistical impasse, we deserve criticism from both economists and statisticians".

3.4 The Mean Square Error Criterion

It has been seen in the preceding section that restrictions on the regression model result in reduction of the variances of the regression estimates, though the restricted estimator will be biased if the restriction is not exactly true. Testing procedures have been devised for rejecting or adopting restrictions on the parameter space in a regression model.

The classical procedure for testing the validity of the restriction $H'\beta = h$, where H is $m \times k$ of rank k, has been the Snedecor F test which can be shown to be uniformly most powerful (U.M.P.).

It is obtained via the statistic

$$z = \frac{SSE(\tilde{\beta}^*) - SSE(\hat{\beta})}{k} \div \frac{SSE(\hat{\beta})}{n - m}$$

where $SSE(\tilde{\beta}^*)$ is the error sum of squares in the least squares regression constrained by the hypothesis and $SSE(\hat{\beta})$ is the error sum of squares in the ordinary least squares (o.ℓ.s.) regression. Under the null hypothesis, $H'\beta = h$, z has the central F distribution with k and $n - m$ degrees of freedom. Consequently, we are able to employ the F test to choose between two sets of estimators, as for example, a variable is dropped when found insignificant using the F test.

A number of disadvantages arise however, from using this test, as Wallace [59] has recently reiterated. Most importantly, the validity of $H'\beta = h$ constitutes an "overstrong" criterion, and even in cases where multicollinearity is severe, it would still seem reasonable to trade some bias for a smaller variance of the estimator. As an alternative which better captures the "notion of tradeoff between bias and variance", Toro-Vizcarrondo and Wallace [56] have proposed a testing procedure based on the Mean Square Error criterion.

The mean square error for an estimator is the estimator's variance plus the square of its bias. Let $\hat{\theta}$ be an $m \times 1$ vector of estimates, then

$$\begin{aligned} MSE_{\hat{\theta}\hat{\theta}} &= E(\hat{\theta}-\theta)(\hat{\theta}-\theta)' \\ &= \Sigma_{\hat{\theta}\hat{\theta}} + (\text{bias } \hat{\theta})(\text{bias } \hat{\theta})' \quad . \end{aligned}$$

Between two alternative $m \times 1$ vector estimators, $\hat{\theta}$ and $\tilde{\theta}$, $\hat{\theta}$ is said to be better in the Mean Square Error (MSE) sense if for any $m \times 1$ vector $d \neq 0$

$$\text{MSE } d'\hat{\theta} \leq \text{MSE } d'\tilde{\theta} .$$

This inequality is equivalent to the requirement

$$E(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)' - E(\hat{\theta}-\theta)(\hat{\theta}-\theta)' = \text{a positive semi-definite matrix} .$$

Thus, in the linear regression framework, we question whether

$$\text{MSE}_{\hat{\beta}\hat{\beta}} - \text{MSE}_{\tilde{\beta}^*\tilde{\beta}^*} \text{ is a positive semi-definite matrix} . \tag{28}$$

If it is so, $\tilde{\beta}^*$ is the better estimator by the MSE criterion.

Toro-Vizcarrondo and Wallace derive the condition for (28) to hold as

$$\frac{(H'\beta-h)'(H'(X'X)^{-1}H)^{-1}(H'\beta-h)}{\sigma^2} \leq 1 . \tag{29}$$

Dividing both sides of (29) by 2 and denoting the left hand side so obtained by λ, (29) can be restated as

$$\lambda \leq \frac{1}{2} .$$

In other words, in testing whether $\tilde{\beta}^*$ is better than $\hat{\beta}$ according to the mean square error criterion, the hypothesis of interest is

$$H_0: \lambda \leq \frac{1}{2} .$$

They further show that the statistic z has a non-central F distribution with parameters k, $n - m$, λ. Under the null hypothesis $H'\beta = h$ or equivalently $\lambda = 0$, z has the central F distribution as stated earlier.

Applying a theorem from Lehmann [34] and making the transformation

$$w = \frac{kz}{n-m} + kz$$

they obtain a U.M.P. test for the MSE criterion

$$H_0: \lambda \leq \frac{1}{2} \quad \text{against} \quad H_1: \lambda > \frac{1}{2}$$

$$\text{Accept } H_0: \text{ if } w^* < w_\alpha$$

$$\text{Reject } H_0; \text{ if } w^* \geq w_\alpha \tag{30}$$

w^* stands for the computed value of w and the critical point w_α is determined by

$$\int_0^{w_\alpha} h_{\frac{1}{2}}(w)dw = 1 - \alpha$$

where $h_\lambda(w)$ is the density function of w which can be easily derived from the non-central density of z as

$$h_\lambda(w) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \frac{\Gamma(\frac{2i+k+n-m}{2})}{\Gamma(\frac{n-m}{2})\Gamma(\frac{2i+k}{2})} w^{\frac{1}{2}(2i+k-2)} (1-w)^{\frac{1}{2}(n-m-2)} \tag{31}$$

$h_\lambda(w)$ can be recognized as the beta distribution for $\lambda = 0$ and the non-central beta distribution for $\lambda > 0$.

Since multicollinearity is closely linked to zero restrictions- dropping a variable or set of variables, (30) can be regarded as a U.M.P. test for "multicollinearity". We delete the set of variable X_{m-r} $(n \times m-r)$ from our partitioned model

$$y = X_r \beta_r + X_{m-r} \beta_{m-r} + u$$

when we consider X_{m-r} to be multicollinear with X_r $(n \times r)$, i.e. when the null hypothesis, $\lambda \leq \frac{1}{2}$ is accepted.

In essence then the Mean Square criterion test takes into account both the bias and the variance, rendering an operational advantage over the standard F test mentioned at the beginning of the discussion.

Two alternative but weaker criteria of the Mean Square Error have been recently developed by Wallace [59]. He refers to them as the First Weak Mean Square Error and the Second Weak Mean Square Error criterion.

According to the 1st weak criterion, $\tilde{\beta}^*$ is better in average squared distance if

$$\text{tr } (MSE_{\underset{\beta}{\sim}\underset{\beta}{*}\underset{\beta}{\sim}\underset{\beta}{*}}) \leq \text{tr } (MSE_{\underset{\beta}{\wedge}\underset{\beta}{\wedge}}) \tag{32}$$

which holds when

$$\lambda \leq \frac{1}{2} \lambda_m \text{ tr } (X'X)^{-1} H (H'(X'X)^{-1} H)^{-1} H' (X'X)^{-1}$$

and λ_m is the smallest eigenvalue of $(X'X)$.

The 2nd weak MSE criterion is a test of the betterness of the restricted over the unrestricted estimator of $X\beta = E(y|X)$. $X\tilde{\beta}^*$ is said to be the better estimator of $E(y|X)$ in weak mean squared error iff

$$E(X\beta-X\tilde{\beta}^*)'(X\beta-X\tilde{\beta}^*) \leq E(X\beta-X\hat{\beta})'(X\beta-X\hat{\beta})$$

or equivalently

$$E(\tilde{\beta}^*-\beta)'X'X(\tilde{\beta}^*-\beta) \leq E(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) \ . \qquad (33)$$

A necessary and sufficient condition for (33) to hold is

$$\lambda \leq \frac{k}{2} \ .$$

To recapitulate the gist of the foregoing discussion, it may simply be stated that average squared distance criteria for linear restrictions in regression yield operational tests more appropriate for deciding the exclusion of variables in the event of multicollinearity.

3.5 Principal Component Estimators

The initial impetus to the use of principal component estimators in situations of multicollinearity was provided by Kendall [29]. Suppose we have a matrix X of n observations on m variables, where the observations are expressed in deviation form from the sample means, the principal components of X are the artificial variables $Z_1, Z_2, \ldots, Z_m$ which are linear combinations of the X_i 's so chosen that the variance of Z_1 is a maximum, the variance of Z_2 is a maximum subject to the condition that Z_2 is orthogonal to Z_1 and so forth. Let a be an m-component column vector such that $a'a = 1$. The variance of Xa

is

$$(Xa)^2 = a'X'Xa \ . \tag{34}$$

In finding a normalized eigenvector $a_1' = (a_{11}, a_{12}, \cdots, a_{1m})$ which maximizes (34), we seek the solution to the equation

$$(X'X - \lambda_1 I) = 0$$

where λ_1 is the largest eigenvalue of $(X'X)$. It can be seen that a_1 is the eigenvector of $(X'X)$ corresponding to the eigenvalue λ_1 . $Z_1 = Xa_1$ then constitutes the first principal component. The second principal component, $Z_2 = Xa_2$, where a_2 is the eigenvector corresponding to the second largest eigenvalue of $(X'X)$, is found by maximizing the objective function

$$\phi_2 = a'X'Xa - \gamma(a'a - 1) - \mu(a'a_1) \ .$$

Proceeding in this manner, we obtain all m principal components of X given by

$$Z = XA \ . \tag{35}$$

Thus it turns out that A is an orthogonal matrix and is composed of normalized eigenvectors a_i corresponding to decreasing eigenvalues λ_i of $(X'X)$. The matrix of eigenvalues

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_m \end{pmatrix}$$

satisfies

$$\Lambda = A'X'XA = Z'Z \ .$$

To explain how principal component analysis resolves multicollinearity, let us suppose that the analysis have been applied to the variables X_i . The regression model can then be written in terms of the components as

$$\begin{aligned} y &= X\beta + u \\ &= ZA'\beta + u \\ &= Z\dot{A} + u \end{aligned}$$

where $\dot{A} = A'\beta$.

In this case, the Gauss Markov theorem is applicable and the least squares estimator $\hat{\dot{A}}$ of $\dot{A}$ is then obtained. We have, from (36)

$$\hat{\beta} = A\hat{\dot{A}} \ . \tag{37}$$

In Kendall's view, a better estimator of β than the ordinary least squares (o.ℓ.s) estimator is afforded by deleting from A those components corresponding to small eigenvalues. The estimator so obtained is referred to as the principal component estimator denoted by b^* . In symbols,

$$b^* = A^*\hat{\dot{A}} \tag{38}$$

where $A^* = A\Delta$ and Δ is a diagonal matrix with δ (a binary m-component vector, each element is either 0 or 1) down the principal diagonal. It can be shown that b^* is distributed as $N(A^*\dot{A}, \sigma^2 A^* \Lambda^{-1} A^{*\prime})$.

The justification for the betterness of b^* over $\hat{\beta}$ is evident from the following: Let $b_i^*(h)$ represent the principal component estimate of β_i obtained by deletion of the component h. Var $\hat{\beta}_i$ then exceeds var $b_i^*(h)$ by the amount $\sigma^2 a_{ih}^2/\lambda_h$ which is necessarily positive.

The desirability of using a principal component estimator rather than an o.ℓ.s. estimator has been further delineated by McCallum [40]. In essence, his proposal entails adopting the Mean Square Error as a criterion for selection of the components. More specifically, the principal component estimate b_i^* of a single parameter is better than the o.ℓ.s. estimate $\hat{\beta}_i$ by the Mean Square Error criterion if

$$\text{MSE } (b_i^*) < \text{MSE } (\hat{\beta}_i) .$$

A component is therefore deleted if its exclusion reduces the mean square error of estimating β_i. Since

$$\begin{aligned} \text{MSE } (b_i^*(h)) &= \text{var } b_i^*(h) + \text{bias } b_i^*(h) \cdot \text{bias } b_i^*(h)' \\ &= \sigma^2 \sum_{j \neq h} \frac{a_{ij}^2}{\lambda_j} + (Eb_i^*(h)-\beta_i)(Eb_i^*(h)-\beta_i)' \end{aligned} \tag{39}$$

and

$$\begin{aligned} \text{bias } b_i^*(h) &= \sum_{j \neq h} a_{ij} A_j - \sum_{j=1}^{m} a_{ij} A_j \\ &= -a_{ih} A_h \\ &= -a_{ih} \sum_{\ell=1}^{m} a_{\ell h} \beta_\ell \qquad \text{(from (36))} \end{aligned} \tag{40}$$

we see that the Mean Square Error criterion per se is not operational. In (39), both the true value of β_i, which we want to estimate, and the σ^2, are unknown. Moreover, (40) tells us that the magnitude of the bias depends upon the true values of $\beta_1, \beta_2, \ldots, \beta_m$. Fortunately, though, this difficulty may be overcome if a priori knowledge of relative parameter magnitude or their estimates is available. Such information will indicate situations in which b_i^* is better than $\hat{\beta}_i$ according to the Mean Square Error criterion.

Farebrother [14] has extended McCallum's analysis to the general criterion of minimizing a weighted sum of elements in the equation

$$MSE(b^*) = E(b^*-\beta)(b^*-\beta)' : \tag{41}$$

The off-diagonal elements of (41)

$$E(b_i^*-\beta_i)(b_j^*-\beta_j)$$

are referred to as the mean product error of b_i^* and b_j^* or $MPE(b_i^*, b_j^*)$.

The minimum weighted mean square error (MWMSE) criterion seeks to minimize the function

$$\sum_{i=1}^{m}\sum_{j=1}^{m} F_{ji} \cdot MPE\ (b_i^*, b_j^*) = \text{tr MSE}\ (b^*) \cdot F \tag{42}$$

where F is an $m \times m$ matrix of fixed weights.

Since it follows from (39) and (40) that

$$\begin{aligned} \text{tr MSE}\ (b^*) \cdot F &= \text{tr var}\ b^* \cdot F + \text{tr}\ (A-A^*)A \cdot A'(A-A^*) \cdot F \\ &= \text{tr var}\ \hat{\beta} \cdot F + \text{tr}\ (A-A^*)(AA'-\sigma^2\Lambda^{-1})(A-A^*)' \cdot F \end{aligned}$$

of which the first term on the right side is constant, Farebrother is able to show that (42) is equivalent to minimizing

$$\text{tr } (AA'-\sigma^2\Lambda^{-1})(A-A^*)'F(A-A^*) \ .$$

The principal component estimator b^* is thus better than the o.ℓ.s. estimator $\hat{\beta}$ by the MWMSE criterion if

$$\text{tr } (AA'-\sigma^2\Lambda^{-1})(A-A^*)'F(A-A^*) < 0 \ . \qquad (43)$$

Deletion of a component is therefore desirable if its exclusion reduces the left side of (43).

The work of McCallum and Farebrother described above has been concerned with finding a better estimate of β . Recently, Mitchell [41] chose instead to focus on obtaining a good estimate of $X\beta$. Adopting the 2nd weak MSE criterion, he suggests minimizing the "average prediction mean square error APMSE" of Xb^* as an estimate of $X\beta$

$$\text{APMSE}(Xb^*) = \frac{1}{n} E(Xb^*-X\beta)'(Xb^*-X\beta)$$

$$= \frac{1}{n} \text{tr MSE } (b^*)X'X$$

which is (42) with $F = \frac{1}{n} X'X$.

As a practical illustration of the method, we use once again our French economy data (Chapter I). On applying a principal component analysis to the sample correlations matrix using CEIGS [8], results shown in Table IV and V are obtained.

It is seen that the first two components account for nearly all of the total variability. Given the small value of λ_3, the contribution of the third component can be neglected.

TABLE IV

Eigenvalues of Correlation Matrix for the French Economy Data

Component	Eigenvalue	Percentage of Variability
1	2.08388	69.46
2	0.91505	30.50
3	0.00107	0.04

We thus have

$$Z_1 = 0.68104X_1 + 0.26960X_2 + 0.68081X_3$$

$$Z_2 = 0.18971X_1 - 0.96297X_2 + 0.19156X_3 .$$

TABLE V

Normalized Eigenvectors for the First Two Components

Variable	1	2
X_1	0.68104	0.18971
X_2	0.26960	-0.96297
X_3	0.68081	0.19156

Since the correlation of y with the X's are respectively 0.98418, 0.26591 and 0.98477,

$$\hat{A}_1 = \frac{1}{2.08288}\,[(0.68104)(0.98418)+(0.26960)(0.26591)+(0.68081)(0.98477)]$$

$$= 0.67777$$

$$\hat{A}_2 = \frac{1}{0.91505}\,[(0.18971)(0.98418)+(-0.96297)(0.26591)+(0.19156)(0.98477)]$$

$$= 0.13036\ .$$

From equation (38),

$$b^* = \begin{pmatrix} 0.68104 & 0.18971 \\ 0.26960 & -0.96297 \\ 0.68081 & 0.19156 \end{pmatrix} \begin{pmatrix} 0.67777 \\ 0.13036 \end{pmatrix}$$

$$= \begin{pmatrix} 0.48632 \\ 0.05720 \\ 0.48640 \end{pmatrix}\ .$$

The regression equation may now be re-expressed in terms of the standardized variables,

$$y = 0.48632\tilde{X}_1 + 0.05720\tilde{X}_2 + 0.48640\tilde{X}_3\ . \qquad (44)$$

In accordance with the theory discussed above, the estimates b_i^* of β_i are biased but have smaller variance then the o.l.s. estimator. (44) is thus the regression equation for our French economy data corrected for multicollinearity.

Our discussion in this section has therefore elucidated the utility of the method of principal component in achieving orthogonization of the regression calculation. In addition, the method has the desirable feature of involving a minimum of assumptions and, given the availability of electronic computers, it is basically easy to apply. Two limitations to the procedure however exist, the first being that it works only for linear models. Secondly, in those situations where a priori information about the grouping of the variables is available, other procedures are more appropriately employed.

3.6 Factor Analysis

In a 1966 paper, Scott [50] proposes applying the well known procedure of factor analysis to resolve multicollinearity in regression analysis. Essentially his method involves adding the correlation coefficients between the dependent variable and the independent variable to the correlation matrix. Applying factor analysis to the augmented matrix, Scott then derives the appropriate regression coefficients from the factors obtained. To understand his procedure, a brief look at the factor model is necessary.

The factor model can be expressed briefly as

$$x = Bf + \varepsilon \tag{45}$$

where

x is an $m \times 1$ vector of m standardized variables.

B is the $m \times p$ matrix of factor loadings, $p < m$.

f is a $p \times 1$ vector of factors.

ε is the $m \times 1$ error vector which is distributed independently of f. Both f and ε have multivariate normal distributions.

$E(\varepsilon) = 0$ and $E(f) = 0$

$E(\varepsilon\varepsilon') = V$, a diagonal matrix

$E(ff') = I$, i.e. the factors are uncorrelated and with unit variance.

It then follows that the covariance matrix of x is given by

$$\Sigma_x = BB' + V . \tag{46}$$

A number of different methods exist for determining the matrix B, including the method of principal factor solution, maximum likelihood method, Whittle least squares method, canonical factor analysis and Joreskog method.

Following determination of the matrix B, the factors can be obtained in at least three different forms

$$(B'B)^{-1}B'x = f \tag{47}$$

$$B'\Sigma_x^{-1}x = f \tag{48}$$

$$(I + B'V^{-1}B)B'V^{-1}x = f . \tag{49}$$

Scott derives from the factor model a stochastic linear equation called factor analysis regression which can be used in place of the least squares regression when multicollinearity is present.

Assuming X_1 is the dependent variable and that $N = B'\sum_x^{-1}$, we have from (48) $f = Nx$. Substituting Nx for f in (45), Scott then derives by simple algebraic manipulation the factor analysis regression equation as

$$X_1 = \frac{\sum_{j=1}^{p} b_{1j}n_{j2}}{1 - \sum_{j=1}^{p} b_{1j}n_{j1}} X_2 + \frac{\sum_{j=1}^{p} b_{1j}n_{j3}}{1 - \sum_{j=1}^{p} b_{1j}n_{j1}} X_3 + \ldots + \frac{\sum_{j=1}^{p} b_{1j}n_{jm}}{1 - \sum_{j=1}^{p} b_{1j}n_{j1}} X_m . \quad (50)$$

Let b_i' be a row vector of the matrix B and n_i be a column vector of matrix N. Equation (50) then assumes a simpler form

$$X_1 = \frac{b_1'n_2}{1 - b_1'n_1} X_2 + \frac{b_1'n_3}{1 - b_1'n_1} X_2 + \ldots + \frac{b_1'n_m}{1 - b_1'n_1} X_m . \quad (51)$$

Putting $W = BB'\sum_x^{-1}$, the factor analysis equation reduces to

$$X_1 = \frac{w_{12}}{1 - w_{11}} X_2 + \frac{w_{13}}{1 - w_{11}} X_3 + \ldots + \frac{w_{1m}}{1 - w_{11}} X_m$$

where w_{ij} is the element of W in the i-th row and j-th column.

In general, any one of the variables may be the dependent variable. Suppose the i-th variable is selected as the dependent variable

$$X_i = \frac{w_{i1}}{1 - w_{ii}} X_1 + \frac{w_{i2}}{1 - w_{ii}} X_2 + \ldots + \frac{w_{im}}{1 - w_{ii}} X_m \qquad (52)$$

where all variables except X_i appear on the right.

If the factor solution given in (47) is employed, the linear model derived by assuming $\tilde{W} = B(B'B)^{-1}B'$ will be

$$X_i = \frac{\tilde{w}_{i1}}{1 - \tilde{w}_{ii}} X_1 + \frac{\tilde{w}_{i2}}{1 - \tilde{w}_{ii}} X_2 + \ldots + \frac{\tilde{w}_{im}}{1 - \tilde{w}_{ii}} X_m$$

where all variables except X_i appear on the right.

On the other hand, if we use (49) and denote $B(I + B'V^{-1}B)^{-1}B'V^{-1}$ by $\hat{W}$, we obtain the linear model

$$X_i = \frac{\hat{w}_{i1}}{1 - \hat{w}_{ii}} X_1 + \frac{\hat{w}_{i2}}{1 - \hat{w}_{ii}} X_2 + \ldots + \frac{\hat{w}_{im}}{1 - \hat{w}_{ii}} X_m .$$

As before, all variables except X_i appear on the right.

To conclude, we note with interest Scott's recommendation that stochastic linear equations derived from factor analysis are especially appropriate for economic data involving high multicollinearity or errors in the variables. The rationale is that the coefficients so obtained are better from the view point of "their economic meaning and theoretical expectation" than those estimated by traditional least squares. Thus, given the availability nowadays of electronic computer for the iterative-type calculation needed, the factor analysis regression may well see more use in econometrics than has hereto occurred.

3.7 Ridge Analysis

In all the preceding procedures for resolving multicollinearity, the estimator of β is the least squares vector $\hat{\beta}$. In 1970, Hoerl and Kennard [24] published an alternative method of estimation known as ridge regression in which a biased estimator $\hat{\beta}^*$ is introduced, namely,

$$\hat{\beta}^* = (\tilde{X}'\tilde{X} + kI)^{-1}\tilde{X}'\tilde{y}$$

$\hat{\beta}^*$ is related to the least squares estimator in the form

$$\hat{\beta}^* = [I + k(X'X)^{-1}]^{-1}\hat{\beta} .$$

In addition, for $k \neq 0$, $\hat{\beta}^{*'}\hat{\beta}^* < \hat{\beta}'\hat{\beta}$, i.e. $\hat{\beta}^*$ is shorter than $\hat{\beta}$.

Basically, the idea of ridge regression is that when a small positive number k is added to the diagonal elements of $\tilde{X}'\tilde{X}$, the instability of the estimates is lowered. This can be seen from the fact that if λ_i is the eigenvalue of $\tilde{X}'\tilde{X}$, then $1/\lambda_i + k$ is the eigenvalue of $[\tilde{X}'\tilde{X} + kI]^{-1}$. More specifically, Hoerl and Kennard have shown that choice of an optimum k can, in fact, reduce the variance and lead to a minimum value of the mean square error of the estimate of β.

The optimal value of k is manifested by a number of simultaneous conditions, namely, the stabilization of the estimates, the disappearance of unreasonably large absolute value of the coefficients, the correction of wrong signs of the coefficients and the reduction of unreasonably large value of the residual sum of

squares. To detect this optimal k, Hoerl and Kennard utilize Ridge-Trace which is a two-dimensional plot of $\hat{\beta}^*(k)$ and the residual sum of squares for the number of values of k in the interval $[0,1]$. In sum, Ridge-Trace reflects the complex interrelationships existing among the non-orthogonal independent variables and the effect of these interrelationships on the estimate of β.

As the authors of this technique have pointed out, Ridge Regression presents two advantages over procedures such as principal components and zero restrictions which do not portray how multi-collinearity is actually causing instability, over-estimations and incorrect signs. In addition, "they can actually amplify the deficiencies of ordinary least squares for non-orthogonal data". We note, of course, the presence of subjectivity in interpretation of the Ridge-Trace to obtain the optimum k.

Most recently, Conniffe and Stone [12] have made some critical comments on Ridge Regression. Their principal contention is that Hoerl and Kennard's proof that the mean square error of $\hat{\beta}^*$ is less than that of $\hat{\beta}$ for certain values of k is valid only if k is assumed known. The ridge procedure however involves the estimation of k. Mayer and Willke [39] confirming this oversight of Hoerl and Kennard, have listed two resultant weaknesses in Ridge Analysis. First, it is not possible to state with absolute certainty that the estimator chosen has smaller total mean square error than the variance of the least squares estimator. Secondly, the moments of $\hat{\beta}^*$ obtained for fixed k are not the moments of the estimator being used.

Conniffe and Stone also argue, quite correctly, that Hoerl and Kennard provide no proof that the appropriate value of k can be recognized by the four criteria discussed previously. Their additional comment is that the second and third criteria entail a lot of prior knowledge which they claim a researcher rarely has. In this regard, it might be countered that in economic situations, at least, theory has developed to a stage where the true nature of the variables' intercorrelation is known. For example, one would expect the coefficient of the rate of change in wages to have positive sign in a price and wage change relationship.

A third point of Conniffe and Stone relates to the stabilization criteria. By showing that even if the $\tilde{X}_i$'s are orthogonal, $\hat{\beta}^*$ values would change more slowly with increasing k, they concluded that the tendency towards stability is not a consequence of the ill-conditioned $(\tilde{X}'\tilde{X})$. The final critical comment refers to the fact that if $(\tilde{X}'\tilde{X})$ is singular, ordinary least squares estimator $\hat{\beta}$ does not exist. However, since $\hat{\beta}^* = (\tilde{X}'\tilde{X} + kI)^{-1}\tilde{X}'\tilde{y}$ and $(\tilde{X}'\tilde{X} + kI)$ is non-singular, $\hat{\beta}^*$ does exist and their values are non-sensical. However, Conniffe and Stone's argument overlooks the relationship existing between $\hat{\beta}$ and $\hat{\beta}^*$, namely

$$\hat{\beta}^* = (I + k(\tilde{X}'\tilde{X})^{-1})^{-1}\hat{\beta}$$

which involves also the inverse of $(\tilde{X}'\tilde{X})$. In sum, taking into account the weaknesses of their critic, Conniffe and Stone's conclusion that "We believe ridge estimators are unlikely to be of practical use to the researcher with data to analyse" would seem slightly over-strong.

3.8 Marquardt Generalized Inverse Estimators

Marquardt [36] has proposed the use of another class of biased estimators termed generalized inverse estimators which share many of the properties of ridge estimators, though they are more relevant when the matrix X is singular.

Let $\tilde{X}'\tilde{X}$ be diagonalized into its matrix of ordered eigenvalues by an orthogonal matrix J such that

$$J'(\tilde{X}'\tilde{X})J = D$$

where

$$J'J = I .$$

Suppose $\tilde{X}'\tilde{X}$ is of rank r so that the last $(m-r)$ elements of D are zero, or nearly so. In the latter case, a criterion for determining the rank r is to preselect ω in the range 10^{-1} to 10^{-7} and then choosing the smallest r satisfying

$$\left| \frac{\sum_{i=m-r}^{m} \lambda_i}{\text{Trace } D} \right| < \omega .$$

To obtain the inverse of $(\tilde{X}'\tilde{X})$, J and D are both partitioned in similar fashion, giving

$$J = (J_r \vdots J_{m-r})$$

$$D = \begin{pmatrix} D_r & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & D_{m-r} \end{pmatrix}.$$

Since D_{m-r} is 0 , D_{m-r}^{-1} equals zero. The inverse $(X'X)$ thus becomes

$$(\tilde{X}'\tilde{X})_r^+ = J_r D_r^{-1} J_r'$$

$$= \sum_{i=1}^{r} \frac{1}{\lambda_i} J_i J_i' \quad . \tag{53}$$

The class of generalized inverse estimators is then defined by

$$\hat{\beta}_r^+ = (\tilde{X}'\tilde{X})_r^+ \tilde{X}'\tilde{y}$$

where $(\tilde{X}'\tilde{X})_r^+$ is as given in (53).

Marquardt indicates that the best r can be selected by examining the size of the variance inflation factor, which is defined as the diagonal element of $(\tilde{X}'\tilde{X})_r^+$ for pre-assigned rank r , $0 < r \leq m$. The criterion suggested is that the maximum variance inflation factor should be "usually larger than 1.0 but certainly not as large as 10" .

In proposing the generalized inverse estimator, Marquardt has emphasized with little reservation its superiority over o.ℓ.s. estimator in non-orthogonal data. Nevertheless, it needs to be pointed out that the same sort of criticisms which have been levied by others against the ridge estimator, apply equally to the generalized inverse estimator. Thus Marquardt criterion for choosing the best r lacks precision, and, more importantly, he has not proved that $\hat{\beta}_r^+$ has a smaller mean square error than the o.ℓ.s. estimator.

3.9 Mayer-Willke Shrunken Estimator

An alternative class of biased estimators, similar to the ridge estimators and labelled shrunken estimators, have been recently proposed by Mayer and Willke [39]. These estimators are defined by

$$\hat{\beta}_\lambda = \lambda(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = \lambda\hat{\beta}, \quad \lambda \in [0,\infty) .$$

If λ is a scalar, $\hat{\beta}_\lambda$ is called a deterministically shrunken estimator. But if λ is a function of $\hat{\beta}'\hat{\beta}$, $\hat{\beta}_\lambda$ is referred to as a stochastically shrunken estimator.

In their paper, Mayer and Willke outline a number of methods by which λ can be selected. On approach involves putting

$$\lambda = [1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}] \tag{54}$$

as the shrinkage factor where

$$s^2 = \tilde{y}'\tilde{y} - \hat{\beta}'(\tilde{X}'\tilde{X})\hat{\beta} .$$

The shrunken estimator is then given by

$$\hat{\beta}_\lambda = [1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}]\hat{\beta} . \tag{55}$$

As Sclove [49] has proved, when the number of independent variables $m \geq 3$, and $0 < \xi < 2(m-2)(n-m+2)^{-1}$, $\hat{\beta}_\lambda$ has smaller minimum weighted mean square error (MWMSE) than $\hat{\beta}$. Indeed, if

$$\xi_0 = (m-2)(n-m+2)^{-1} ,$$

then

$$\text{MWMSE}(\hat{\beta}_{\lambda(\xi_0)}) = \min_{\xi} \text{MWMSE}(\hat{\beta}_{\lambda(\xi)}) .$$

Thus the class of stochastically shrunken estimator with λ as defined in (54) is superior to the ridge estimators or deterministically shrunken estimators, since a value of λ can be determined which will guarantee a better estimator of β than the ordinary least squares estimator $\hat{\beta}$, "betterness" being in the sense of the Minimum Weighted Mean Square Error criterion. It must be stressed, however, that shrunken estimators with other values of λ face the same problems as ridge estimators because they involve the estimation of λ.

3.10 Multicollinearity in Two-stage Least-squares

The techniques discussed thus far are designed to resolve multicollinearity when it occurs in ordinary least squares estimation. As Klein and Nakamura [31] have shown, two-stage least-squares estimations are even more sensitive to the presence of multicollinearity, and a remedy for such situations has been devised by Kloek and Mennes [32]. To understand their procedure, a brief discussion of two-stage least-squares will be necessary.

In brief, the model concerned is

$$y = Y_1\beta + X_1\alpha + u$$

where y is an $n \times 1$ vector of observations on the "dependent" variables.

Y_1 is an $n \times \ell$ matrix of the other endogeneous variables present in the equation. X_1 is an $n \times k$ matrix of observations on the predetermined variables appearing in the equation. u is an $n \times 1$ vector of disturbances.

Basically, two-stage least-squares estimation involves replacing Y_1 by their least squares estimator $\hat{Y}_1$, where

$$\hat{Y}_1 = X(X'X)^{-1}X'Y_1$$

$$X = [X_1 \quad X_2]$$

and X_2 is an $n \times (K-k)$ matrix of predetermined variables not appearing in the equation. Next y is regressed on $\hat{Y}_1$ and X_1 to obtain two-stage least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ as

$$\begin{bmatrix} Y_1'Y_1 - V_1'V_1 & Y_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{bmatrix} (Y_1 - V_1)'y \\ X_1'y \end{bmatrix}$$

where V_1 is the $n \times \ell$ matrix of residuals from the least squares regression of Y_1 on X.

As is well known, difficulties arise when the number of predetermined variables exceeds the number of observations, or when the number of degrees of freedom for the regressions is unsatisfactorily small. An attempted solution to these difficulties consists of replacing X_2 by a small number of principal components. Unfortunately, as it often turns out, multicollinearity may exist between one or more of the components with some of the variables in X_1.

To resolve this impasse, Kloek and Mennes have suggested a number of alternative methods, each beginning with normalization of all predetermined variables.

One alternative involves utilization of the principal components of the residual when X_2 is regressed on X_1, the residual being

$$E = X_2 - X_1(X_1'X_1)^{-1}X_1'X_2$$

The principal components used are then given by

$$e_j = Eo_j$$

where o_j is the eigenvector of $E'E$ corresponding to the eigenvalue λ_j, $j = 1,\ldots,K-k$.

Multicollinearity is therefore avoided as the following argument demonstrates.

$$\begin{aligned} X_1'e_j &= X_1'Eo_j \\ &= X_1'[X_2 - X_1(X_1'X_1)^{-1}X_1'X_2]o_j \\ &= [X_1'X_2 - X_1'X_2]o_j \\ &= 0 \qquad\qquad j = 1,\ldots,K-k . \end{aligned}$$

A second alternative involves selecting those components with the greatest θ_j, defined as

$$\theta_j = \lambda_j''(1 - R_j^2) \qquad j = 1,\ldots,K-k$$

where R_j is the multiple correlation coefficient when P_j, the j-th principal component of X_2, is regressed on X_1 and λ_j'' is the j-th eigenvalue of $(X_2'X_2)$.

Multicollinearity is resolved since

$$R_j^2 = \frac{P_j'X_1(X_1'X_1)^{-1}X_1'P_j}{P_j'P_j}$$

by definition and $P_j'P_j = \lambda_j''$, so that

$$\lambda_j''(1-R_j^2) = P_j'[I-X_1(X_1'X_1)^{-1}X_1']P_j \ . \qquad (56)$$

The right side of (56) is the residual sum of squares when P_j is regressed on X_1. In other words, the components chosen are those which are least correlated with X_1.

CHAPTER IV

AN EMPIRICAL STUDY OF MULTICOLLINEARITY

Our discussion in the preceding chapters has presented an overview of the problems, detection and correction of multicollinearity in regression analysis. To illustrate the practical implications of the theory reviewed, we have empirically investigated multicollinearity in economic data related to inflation, a problem of considerable current interest and significance. In essence, we will attempt to apply the Farrar-Glauber techniques to the most recent data available concerning price changes in relation to the rate of change in wages and certain other contributing factors. This is followed by the construction of Hoerl and Kennard's Ridge Trace to visually demonstrate the harmful effects of multicollinearity in our sample of data. Finally, Mayer and Willke's shrunken estimator is calculated to remedy the detected multicollinearity.

4.1 Description of the Model

The economic model we employ is that found in the Special Study No. 5 published by the Economic Council of Canada in 1965. This study estimated the price change equation by fitting regressions to quarterly data over the period 1949 : 1 to 1965 : 2 . For the purposes of our investigation, the same relationship is utilized for quarterly data which has been collected for the period 1959 : 1 to 1972 : 4 .

We have the model

$$\dot{P}_t = \beta_0 + \beta_1 \dot{W}_t + \beta_2 \dot{F}_t + \beta_3 \dot{P}_{US_t} + \beta_4 \dot{P}_{t-1} + \beta_5 \dot{P}_{t-2}$$

where

$\dot{P}_t = \frac{P_t - P_{t-4}}{P_{t-4}} \cdot 100$ = percentage change in the Consumer Price Index between the current quarter and the same quarter of the preceding year. (DBS: 62-002).

$\dot{W}_t = \frac{W_t - W_{t-4}}{W_{t-4}} \cdot 100$ = percentage change in average hourly earning of production workers in manufacturing (DBS: 72-204).

$\dot{F}_t = \frac{F_t - F_{t-4}}{F_{t-4}} \cdot 100$ = percentage change in the implicit deflator for imports of goods and services in the National Accounts (DBS: 13-001).

$\dot{P}_{US_t} = \frac{P_{US_t} - P_{US_{t-4}}}{P_{US_{t-4}}} \cdot 100$ = percentage rate of change in the U.S. Consumer Price Index (Labour Review, U.S. Department of Labour).

$\dot{P}_{t-1}$ = the value of $\dot{P}_t$ in the immediate preceding quarter

$\dot{P}_{t-2} = \dot{P}_t$ lagged two quarters.

4.2 Application of Farrar-Glauber technique

Using least-squares analysis computer program MLREGR, the estimated relationship is obtained as

$$\dot{P}_t = -0.44201 + 0.24887\dot{W}_t + 0.11055\dot{F}_t$$
$$(2.99345) \quad (1.91801)$$
$$-0.19881\dot{P}_{US_t} + 0.87283\dot{P}_{t-1} - 0.09112\dot{P}_{t-2} \; . \tag{57}$$
$$(-1.83583) \quad (6.22947) \quad (-0.59336)$$

The squared multiple correlation coefficient

$$R^2 = 0.87400 \; .$$

The matrix of simple correlation coefficient between the independent variables is

$$C = \begin{pmatrix} 1.00000 & -0.07353 & 0.89543 & 0.81324 & 0.83302 \\ -0.07352 & 1.00000 & 0.05934 & 0.02715 & -0.04068 \\ 0.89543 & 0.05934 & 1.00000 & 0.78036 & 0.82513 \\ 0.81324 & 0.02715 & 0.78036 & 1.00000 & 0.91583 \\ 0.83302 & -0.04068 & 0.82513 & 0.91583 & 1.00000 \end{pmatrix} .$$

Since $|C| = 0.00724$, substantial multicollinearity exists among the independent variables.

$\chi^2_{|C|}$ is calculated and is equal to

$$-[52 - 1 - \tfrac{1}{6}(15)](-4.9282) = 239.018 \; .$$

By regressing each independent variable on the remaining ones, we obtain the values of the multiple correlation coefficient $R^2_{X_i}$ and the associated F statistic as follows:

	$\dot{W}_t$	$\dot{F}_t$	$\dot{P}_{US_t}$	$\dot{P}_{t-1}$	$\dot{P}_{t-2}$
$R^2_{X_i}$	0.850	0.127	0.839	0.854	0.876
F	66.583	1.709	61.231	68.295	83.008

The coefficient of partial correlation between pairs of independent variables and associated t-ratio are calculated and shown in Table VI.

TABLE VI

Partial Correlation Coefficient c_{ij}. and Associated t_{ij}. between Pair of Variables with $R^2_{X_i}$ on Diagonal

	$\dot{W}_t$	$\dot{F}_t$	$\dot{P}_{US_t}$	$\dot{P}_{t-1}$	$\dot{P}_{t-2}$
$\dot{W}_t$	0.850	-0.286	0.691	0.254	0.039
$\dot{F}_t$	-2.046	0.127	0.301	0.207	-0.192
$\dot{P}_{US_t}$	6.553	2.269	0.839	-0.116	0.303
$\dot{P}_{t-1}$	1.800	1.483	-0.806	0.854	0.739
$\dot{P}_{t-2}$	0.268	1.367	2.287	11.162	0.876

In the estimated relationship (57), the coefficient of the variable $\dot{P}_{US_t}$ is observed to have negative sign. This result seems unlikely as one could reasonably expect an upward drift in U.S. prices to be accompanied by a similar drift in Canadian prices. One might suspect multicollinearity is the cause of this phenomenon from the small value of $|C|$.

The squared multiple correlation coefficient R^2 indicates that 87% of the total variation in the Consumer Price Index can be explained by the regression equation. From the F values of the independent variables, one may deduce that $\dot{W}_t$, $\dot{P}_{US_t}$, $\dot{P}_{t-1}$ and $\dot{P}_{t-2}$ are affected by multicollinearity. Indeed, Table VI shows that a linkage exists between $\dot{W}_t$ and $\dot{P}_{US_t}$ and in another instance between $\dot{P}_{t-1}$ and $\dot{P}_{t-2}$.

4.3 Ridge Analysis of the Data

Figures 2 and 3 represent the Ridge Trace that have been obtained by applying Ridge Regression to our set of economic data.

Apparent from the Ridge Trace constructed are the following results:

(1) over-estimation of the coefficients of all the variables when using the least squares estimator is clearly evident.

(2) it is seen that when $k = 0$, the coefficients of the variables $\dot{P}_{t-2}$ and $\dot{P}_{US_t}$ have negative signs which move quickly to zero upon

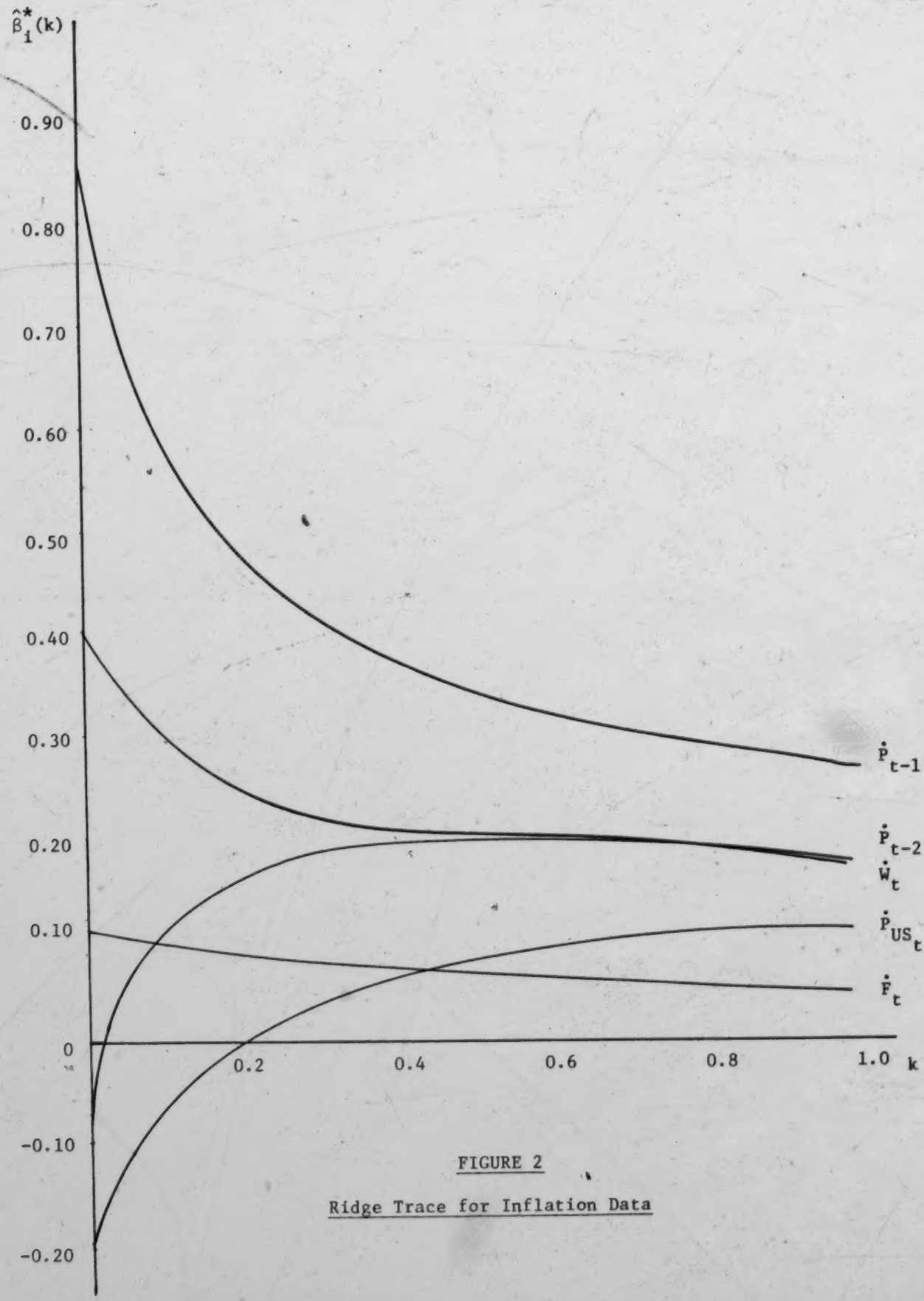

FIGURE 2

Ridge Trace for Inflation Data

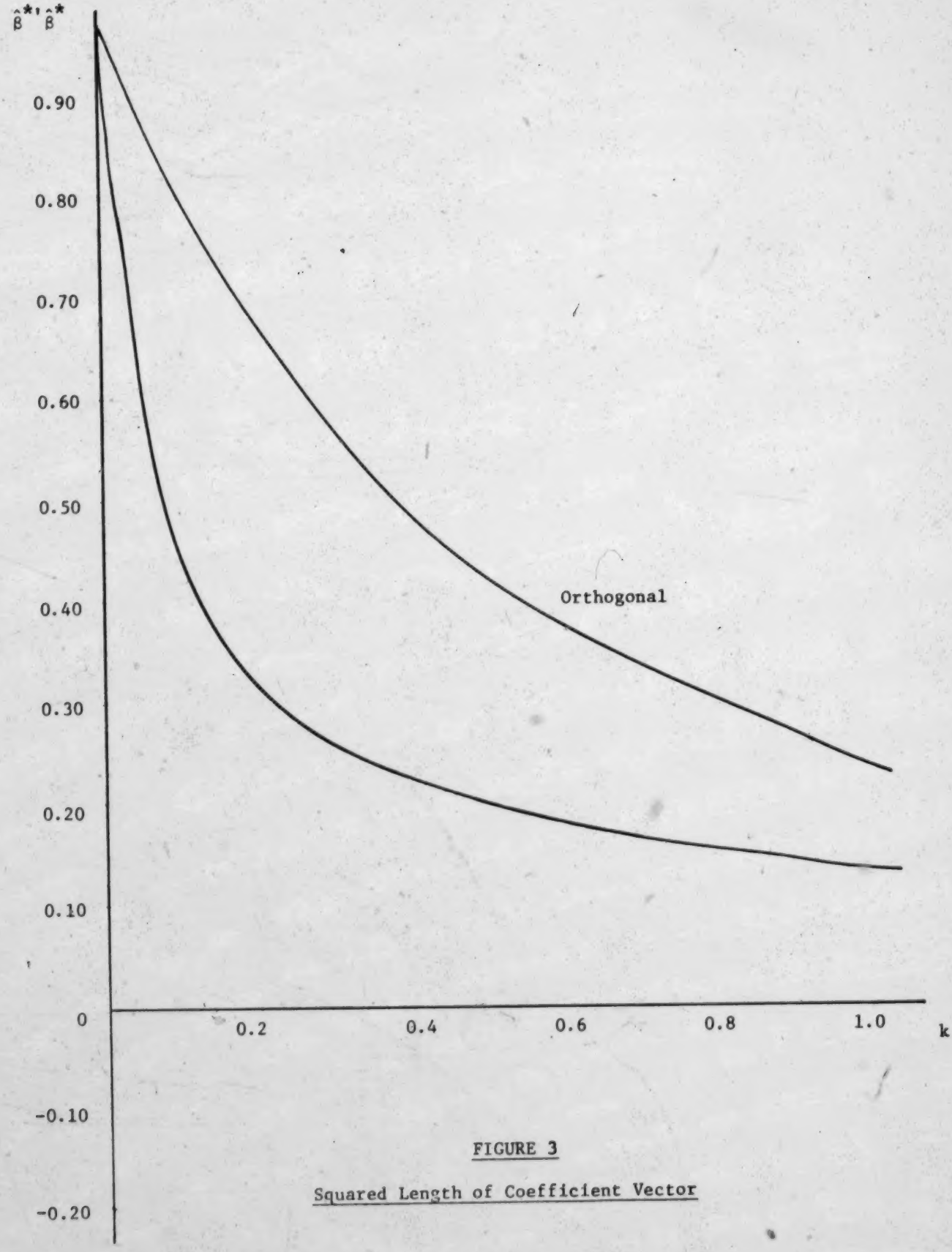

FIGURE 3

Squared Length of Coefficient Vector

the addition of $k > 0$ and subsequently become positive. From this, we deduce that these two coefficients have the wrong signs in the original estimated relationship.

(3) instability characterizes the coefficients of variables $\dot{W}_t$, $\dot{F}_t$, $\dot{P}_{t-1}$ and $\dot{P}_{t-2}$, indicating the presence of multicollinearity among these variables. We note that the same result is obtained using Farrar-Glauber's technique.

(4) the stabilization of the system is observed to occur at a value of k in the interval $(0.5, 0.7)$. Coefficients obtained by employing such a value of k, according to Hoerl and Kennard, affords more stable prediction than the least squares estimator.

4.4 Calculation of Mayer-Willke Shrunken Estimator

Following the procedure of Mayer and Willke, the shrunken estimator for our set of data is calculated (with $\xi = (m-2)(n-m+2)^{-1}$)

$$\hat{\beta}_{\lambda(\xi)} = \begin{pmatrix} 0.40845 \\ 0.10838 \\ -0.24158 \\ 0.86099 \\ -0.08891 \end{pmatrix}.$$

It is observed that the shrunken estimator obtained does not correspond to the ridge estimator calculated using a k value in the range $(0.5, 0.7)$. As we recall, the shrunken estimator has

been shown by Sclove to have smaller minimum weighted mean square error than the least squares estimator. On the other hand, owing to lack of rigorous proof, the superiority of the ridge estimator over least squares estimator is still an issue under debate. Pending the resolution of this controversy, it would seem reasonable therefore to employ the shrunken estimator rather than the ridge estimator as a remedy for multicollinearity in our set of data.

REFERENCES

[1] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York, 1958.

[2] Bartlett, M.S., "Test of Significance in Factor Analysis", British Journal of Psychology, Statistical Section, 3 (1950), 77-85.

[3] Bauer, F.L., "Elimination with Weighted Row Combinations for Solving Linear Equations and Least Squares Problems", Num. Math., 7 (1965), 338-352.

[4] Björck, Å., "Iterative Refinement of Linear Least Solutions I", BIT 7 (1967), 257-278.

[5] Björck, Å., "Solving Linear Least Squares Problems by Gram-Schmidt Orthogonization", BIT 7 (1967), 1-21.

[6] Bose, R.C., Unpublished Lecture Notes on Analysis of Variance, University of North Caroline at Chapel Hill, Chapel Hill, 1959.

[7] Businger, P. and G.H. Golub, "Linear Least Squares Solution by Householder Transformations", Num. Math., 7 (1965), 269-276.

[8] CEIGS Program to find eigenvalues and eigenvectors of a real symmetric matrix, Application Subroutine written by Computing Services, University of Alberta.

[9] Chipman, J.S., "On Least Squares with Insufficient Observations", JASA 59 (1964), 1078-1111.

[10] Chipman, J.S., "Specification Problems in Regression Analysis", Proc. Symposium on Theory and Application of Generalized Inverses of Matrices, Mathematics Series No. 4, Texas Technological College, Lubbock, Texas, 1968.

[11] Conniffe, D. and J. Stone, "A Critical View of Ridge Regression", The Statistician, 22 (1973), 181-187.

[12] Crout, P.D., "A Short Method of Evaluating Determinants and Solving Sets of Linear Equations with Real or Computer Coefficients", Trans. Amer. Inst. Elect. Eng., 60 (1941), 1235-1240.

[13] Dwyer, P.S., "Generalizations of a Gaussian Theorem", Ann. Math. Statist., 29 (1958), 106-117.

[14] Farebrother, R.W., "Principal Component Estimators and Minimum Mean Square Error Criteria in Regression Analysis", The Review of Economics and Statistics, 54 (1972), 332-336.

[15] Farrar, D.E. and R.R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited", The Review of Economics and Statistics, 54 (1972), 332-336.

[16] Feldstein, M.S., "Multicollinearity and the Mean Square Error of Alternative Estimators", Econometrica, 41 (1973), 337-345.

[17] Fox, C., Intermediate Economic Statistics, John Wiley and Sons, Inc., New York, 1968.

[18] Frisch, R., Statistical Confluence Analysis by Means of Complete Regression Systems, Oslo, 1934.

[19] Goldberger, A.S., Econometric Theory, John Wiley and Sons, Inc., New York, 1964.

[20] Golub, G.H. and J.H. Wilkinson, "Note on the Iterative Refinement of Least Squares Solution", Num. Math., 9 (1966), 139-148.

[21] Gupta, R.P., "A Note on Multicollinearity and Imprecise Estimation", Statistiche Heft 14, 1 (1973), 84-87.

[22] Haitovsky, Y., "Multicollinearity in Regression Analysis: Comment", The Review of Economics and Statistics, 51 (1969), 486-489.

[23] Haitovsky, Y., "On the Correlations between Estimated Parameters in Linear Regression" (forthcoming).

[24] Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, 12 (1970), 55-67.

[25] Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems", Technometrics, 12 (1970), 69-83.

[26] Householder, A.S., "Unitary Triangularization of a Nonsymmetric Matrix", J. Assoc. Comput. Mach., 5 (1958), 339-342.

[27] Hsu, P.L., "On the Problem of Rank and the Limiting Distribution of Fisher's Test Function", Annals of Eugenics, 11 (1941), 39 ff.

[28] Johnston, J., Econometric Methods, McGraw-Hill Book Company, Inc., New York, 1972.

[29] Kendall, M.G., A course in Multivariate Analysis, Charles Griffin and Company Ltd., London, 1957.

[30] Kirchdorfer, H., "Möglichkeiten der Analyse und der Einschränkung der Wirkung von Multikollinearität bei linearer Mehrfachregression", Statistische Paxis, 26 (1971/2), 89-93.

[31] Klein, L.R. and M. Nakamura, "Singularity in the Equation Systems of Econometrics: Some Aspects of the Problem of Multicollinearity", International Economic Review, 3 (1962), 274-299.

[32] Kloek, T. and L.B.M. Mennes, "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables", Econometrica, 28 (1960), 45-61.

[33] Kuh, E. and J.R. Meyer, "How Extaneous are Extraneous Estimates?" The Review of Economics and Statistics, 39 (1957), 380-393.

[34] Lehmann, E.L. Testing Statistical Hypotheses, John Wiley and Sons, Inc., New York, 1964.

[35] Malinvaud, E., Statistical Methods of Econometrics, North-Holland, Amsterdam, 1971.

[36] Marquardt, D.W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation", Technometrics, 12 (1970), 591-612.

[37] Martin, R.S., Peters, G. and J.H. Wilkinson, "Iterative Refinement of the Solution of a Positive Definite System of Equations", Num. Math., 8 (1966), 203-216.

[38] Martin, R.S., Peters, G. and J.H. Wilkinson, "Symmetric Decomposition of a Positive Definite Matrix", Num. Math., 7 (1965), 362-383.

[39] Mayer, L.S. and T.A. Willke, "On Biased Estimation in Linear Models", Technometrics, 15 (1973), 497-508.

[40] McCallum, B.T., "Artificial Orthogonization in Regression Analysis", The Review of Economics and Statistics, 53 (1971), 110-113.

[41] Mitchell, B.M., "Estimation of Large Economietric Models by Principal Components and Instrumental Variable Methods", Technical Report No. 28 (Economic Series), (Stanford, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1970).

[42] MLREGR Multiple Linear Regression. IBM Scientific Subroutine Package, written in FORTRAN.

[43] Penrose, R., "A Generalized Inverse for Matrices", Proc. of the Camb. Phil. Soc., 51 (1955), 406-413.

[44] Plackett, R.L., " Historical Note on the Method of Least Squares", Biometrika, 36 (1949), 458-460.

[45] Plackett, R.L., "Some Theorems in Least Squares", Biometrika, 37 (1950), 149-157.

[46] Raduchel, W.J., "Multicollinearity Once Again", Harvard Institute of Economic Research, Discussion Paper #205, (1971), Harvard University.

[47] Rao, C.R., Linear Statistical Inference and Its Applications, John Wiley and Sons, Inc., New York, 1965.

[48] Rao, C.R., "Calculus of Generalized Inverses of Matrices, Part I: General Theory", Sankhyā, Ser. A, 29 (1967), 317-342.

[49] Sclove, S.L., "Improved Estimators for Coefficients in Linear REgression", J. Amer. Statist. Assoc., 63 (1968), 597-606.

[50] Scott, J.T., "Factor Analysis and Regression", Econometrica, 34 (1966), 552-562.

[51] Silvey, S.E., "Multicollinearity and Imprecise Estimation", J. Roy. Statist. Soc., Ser. B, 31 (1969), 539-552.

[52] Theil, H., Economic Forecasts and Policy, second revised edition, North-Holland, Amsterdam, 1961.

[53] Theil, H. and A.S. Goldberger, "On Pure and Mixed Statistical Estimation in Economics", International Economic Review, 2 (1961), 65-78.

[54] Tintner, G., "A Note on Rank, Multicollinearity and Multiple Regression", Ann. Math. Statist., 16 (1945), 304-307.

[55] Tintner, G., Econometrics, JOhn Wiley and Sons, Inc., New York, 1958.

[56] Toro, C.E. and T.D. Wallace, "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression", JASA 63 (1968), 558-571.

[57] Toro, C.E., "Multicollinearity and the Mean Square Error Criterion in Multiple Regression : A Test and Some Sequential Estimator Comparisons", Unpublished Ph.D. thesis, Department of Experimental Statistics, North Caroline State University at Raleigh, 1968.

[58] Von Neumann, J. and H.H. Goldstine, "Numerical Inverting of Matrices of Higher Order", Bull. Amer. Math. Soc., 53 (1947), 1021-1099.

[59] Wallace, T.D., "Weaker Criteria and Tests for Linear Restrictions in Regression", Econometrica, 40 (1972), 689-697.

[60] Wilks, S., "Certain Generalizations in the Analysis of Variance", Biometrika, 24 (1932), 471-494.

[61] Wilkinson, J.H., Rounding Errors in Algebraic Processes, London: Her Majesty's Stationary Office, Prentice-Hall, Eaglewood Cliffs, N.J.,1963.

APPENDIX I

INVERTING ILL-CONDITIONED MATRICES

A well known difficulty in solving least-squares equation stems from the need to invert X, matrices which are often ill-conditioned. To circumvent this difficulty, a common procedure lies in the following process of iterative refinement proposed by Wilkinson [61]. The sequence of vectors $\hat{\beta}^{(s)}$, $s = 0,1,2,..$ defined by

$$\hat{\beta}^{(0)} = 0, \quad r^{(s)} = y - X\hat{\beta}^{(s)}$$

$$\delta\hat{\beta}^{(s)} = X'r^{(s)}, \quad \hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + \delta\hat{\beta}^{(s)}$$

is computed. In the computation of the residual $r^{(s)}$, double precision accumulation of inner products is employed. All other steps are carried out with single precision.

Recent years have seen the development of various algorithms aimed at obtaining more accurate solutions. Some of the more successful methods are those of Businger [7], Martin, Peters and Wilkinson [37, 38], Golub [4, 20], Bauer [3], Björck [4].

Businger and Golub's procedure employs orthogonal Householder transformation. Since length is invariant under orthogonal transformation,

$$\|y-X\hat{\beta}\|_2 = \|Qy-QX\hat{\beta}\|_2$$

the least squares problem reduces to that of minimizing $\|Qy-QX\hat{\beta}\|_2$. Q is chosen in such a way that

$$QX = \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} \, n \times k \\ \} \, n \times (m-k) \end{matrix} \qquad \text{(I-1)}$$

where R is an upper triangular matrix. The decomposition in (I-1) can be accomplished efficiently by the Householder transformation [26] and clearly,

$$\hat{\beta} = R^{-1}\widetilde{Qy}$$

where $\widetilde{Qy}$ is the first k components of Qy .

Once an initial solution has been obtained, it may be improved to considerable accuracy by the process of iterative refinement. Iteration is continued as long as improved estimates of β can be obtained. The iterative technique should be used only if the initial approximation is sufficiently accurate, otherwise the iteration will not converge.

The method of Martin, Peters and Wilkinson decomposes the symmetric, positive definite matrix X into LL' , where L is a non-singular lower triangular matrix. The elements of L are obtained by the Cholesky decomposition [12] and then used to solve the least squares solution. Since

$$X\hat{\beta} = LL'\hat{\beta} = Lv = y$$

we have

$$v_i = (y_i - \sum_{k=1}^{i-1} \ell_{ik} v_k)/\ell_{ii}, \qquad i = 1,\ldots,n$$

$$\hat{\beta}_i = (v_i - \sum_{k=i+1}^{n} \ell_{ki} \hat{\beta}_k)/\ell_{ii}, \qquad i = 1,\ldots,n .$$

In each iteration of the refinement process, $\hat{\beta}^{(s)}$ is improved by a correction $\delta\hat{\beta}^{(s)}$ that is determined using the computed LL' factorization. Ground rules for iteration are again as laid down earlier.

Bauer has formulated an ALGOL procedure in which X is decomposed into GDB, where G consists of orthogonal non-zero columns, $D = (G'G)^{-1}$, and B is upper triangular. The condition $G'(y-X\hat{\beta}) = 0$ yields the triangular system

$$B\hat{\beta} = G'y$$

which is then solved by back-substitution.

The procedure devised by Björck requires decomposing X into

$$X = VC$$

where C is unit upper triangular, and $V'V$ is diagonal. To accomplish the decomposition, Björck uses a modification of the Gram-Schmidt orthogonization process. This differs from the classical process in that the elements of C are computed one row instead of one column at a time. Once the decomposition is realized, the least

squares solution is given by

$$\hat{\beta} = C^{-1}V^{-1}y \ .$$

Iterative refinement is then carried out in the usual way.

Björck has also proposed utilizing the fact that the residual r is orthogonal to the columns of X. His procedure therefore considers the augmented system

$$\begin{pmatrix} I & X \\ X' & 0 \end{pmatrix} \begin{pmatrix} r \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \ .$$

In his three-stage iterative refinement procedure, Björck begins by computing the residuals

$$\begin{pmatrix} f^{(s)} \\ g^{(s)} \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} I & X \\ X' & 0 \end{pmatrix} \begin{pmatrix} r^{(s)} \\ (s) \end{pmatrix}$$

with double precision accumulation of inner products. This is followed by obtaining the corrections $\delta r^{(s)}$ and $\delta\hat{\beta}^{(s)}$ from

$$\begin{pmatrix} I & X \\ X' & 0 \end{pmatrix} \begin{pmatrix} \delta r^{(s)} \\ \delta\hat{\beta}^{(s)} \end{pmatrix} = \begin{pmatrix} f^{(s)} \\ g^{(s)} \end{pmatrix}$$

which can be obtained by the Householder's or modified Gram-Schmidt method. The corrections are then added in the final step,

$$\begin{pmatrix} r^{(s+1)} \\ \hat{\beta}^{(s+1)} \end{pmatrix} = \begin{pmatrix} r^{(s)} \\ \hat{\beta}^{(s)} \end{pmatrix} + \begin{pmatrix} \delta r^{(s)} \\ \delta\hat{\beta}^{(s)} \end{pmatrix} \ .$$

In conclusion, it needs only be said that the above procedures attain their objective highly satisfactorily. Either inversion of the ill-conditioned matrix is achieved to working accuracy, or the system found too ill-conditioned to be solved without working to higher precision. No attempt, however, has yet been made to compare the procedures with respect to computer time required, applicability, storage requirements or program output. A line of numerical analysis research may well be fruitfully pursued towards such a comparison.

APPENDIX II

PROGRAM A

```
C  PROGRAM TO CALCULATE MATRIX U OF KIRCHDORFER PROCEDURE
   DIMENSION XI2(20),DI1(20,XI3(20),EI2(20),SQR(20),DI2(20),EI3(20),
  /EI1(20),XI1(20)
   READ(5,11) L
11 FORMAT(I2)
   DO 10 II=1,L
   READ(5,20) XI1(II),XI2(II),XI3(II)
20 FORMAT(3F6.3)
10 CONTINUE
   U00=SQRT(1.*L)
   DI0=1./U00
   CALL SUMN (DI0,XI1,L,U01)
   CALL SUMN (DI0,XI2,L,U02)
   CALL SUMN (DI0,XI3,L,U03)
   DO 2 I2=1,L
 2 EI1(I2)=XI1(I2)-U01*DI0
   CALL SUMM (EI1,EI1,L,SSS)
   U11=SQRT(SSS)
   DO 4 I4=1,L
 4 DI1(I4)=EI1(I4)/U11
   CALL SUMM (DI1,XI2,L,U12)
   CALL SUMM (DI1,XI3,L,U13)
   SUM=0.
   DO 30 I=1,L
   EI2(I)=XI2(I)-U02*DI0-U12*DI1(I)
   SQR(I)=EI2(I)**2
30 SUM=SUM+SQR(I)
   U22=SQRT(SUM)
   DO 40 J=1,L
40 DI2(J)=EI2(J)/U22
   CALL SUMM (DI2,XI3,L,U23)
   SS=0.
   DO 60 M=1,L
   EI3(M)=XI3(M)-U03*DI0-U13*DI1(M)-U23*DI2(M)
60 SS=SS+EI3(M)**2
   U33=SQRT(SS)
   WRITE(6,123)
123 FORMAT('1',9X,'EI1',13X,'DI1',12X,'EI2',12X,'DI2',12X,'EI3'//)
   DO 70 N=1,L
   WRITE(6,99) EI1(N),DI1(N),EI2(N),DI2(N),EI3(N)
99 FORMAT(' ',5F15.5)
70 CONTINUE
   U10=0.
   U20=0.
   U21=0.
```

```
      U30=0.
      U31=0.
      U32=0.
      WRITE(6,999)
  999 FORMAT('1','   MATRIX U IS  '//)
      WRITE(6,991)
      WRITE(6,991) U00,U01,U02,U03
  991 FORMAT(' ',4F10.5//)
      WRITE(6,992) U10,U11,U12,U13
  992 FORMAT(' ',4F10.5//)
      WRITE(6,993) U20,U21,U22,U23
  993 FORMAT(' ', 4F10.5//)
      WRITE(6,994) U30,U31,U32,U33
  994 FORMAT(' ',4F10.5//)
      WRITE(6,1234)
 1234 FORMAT('1')
      STOP
      END

      SUBROUTINE SUMM (B,X,L,S)
      DIMENSION B(20),X(20)
      S=0.
      DO 1 I=1,L
    1 S=S+B(I)*X(I)
      RETURN
      END

      SUBROUTINE SUMN (BB,X,L,S)
      DIMENSION X(20)
      S=0.
      DO 1 I=1,L
    1 S=S+BB*X(I)
      RETURN
      END
```